RESEARCH ARTICLE • OPEN ACCESS

An Integrated K-Means++–Davies–Bouldin Index Approach for Educational Resource-Based District Clustering: A Case Study of Districts in Surabaya

Hendrik Subaekti, Lutfi Hakim, Hani Khaulasari, and Dian Yuliati

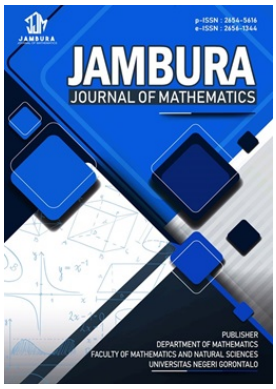
JAMBURA JOURNAL OF MATHEMATICS

Volume 8, Issue 1, Pages 111–120, February 2026

Received 18 November 2025, Revised 11 February 2026, Accepted 26 February 2026, Published 28 February 2026

To Cite this Article : H. Subaekti, L. Hakim, H. Khaulasari, and D. Yuliati,"An Integrated K-Means++–Davies–Bouldin Index Approach for Educational Resource-Based District Clustering: A Case Study of Districts in Surabaya ", *Jambura J. Math*, vol. 8, no. 1, pp. 111–120, 2026, *https://doi.org/10.37905/jjom.v8i1.35412*

JOURNAL INFO • JAMBURA JOURNAL OF MATHEMATICS



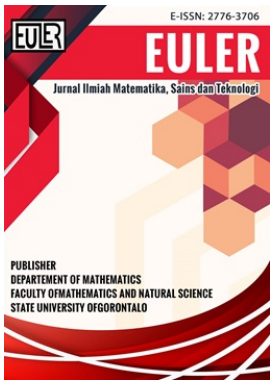
Homepage	:	http://ejurnal.ung.ac.id/index.php/jjom/index
Journal Abbreviation	:	Jambura J. Math.
Frequency	:	Biannual (February and August)
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/jjom
Online ISSN	:	2656-1344
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjom/oai
Google Scholar ID	:	iWLjgaUAAAAJ
Email	:	info.jjom@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

Jambura Journal of Biomathematics

Jambura Journal of Mathematics Education

Jambura Journal of Probability and Statistics



EULER : Jurnal Ilmiah Matematika, Sains, dan Teknologi

Jambura Journal of Mathematics, Volume 8, Issue 1, Pages 111–120, February 2026 https://doi.org/10.37905/jjom.v8i1.35412

Research Article

Check for updates

An Integrated K-Means++–Davies–Bouldin Index Approach for Educational Resource-Based District Clustering: A Case Study of Districts in Surabaya

Hendrik Subaekti[1], **Lutfi Hakim**[1], **Hani Khaulasari**[1,*], **Dian Yuliati**[1]

[1]*Mathematics Department, Universitas Islam Negeri Sunan Ampel Surabaya, Surabaya 60237, Indonesia*

ARTICLE HISTORY
Received 18 November 2025
Revised 11 February 2026
Accepted 26 February 2026
Published 28 February 2026

KEYWORDS
Education
Educational Resources
District Clustering
K-Means++
Davies-Bouldin Index

ABSTRACT. *Equitable distribution of educational resources is an important prerequisite to ensure that all communities benefit from human resource development. Access to education through the availability of schools and teachers at every level, plays a role in reducing the gap between regions. This study aims to group educational resources at the elementary and junior high school levels in 31 sub-districts of Surabaya City and evaluate the quality of grouping using the Davies–Bouldin Index (DBI). The analysis was carried out using secondary data from the Surabaya City Education Office which included the number of schools, teachers, and students based on education level in each sub-district. The clustering method used is K-Means++, which improves the centroid initialization process to produce more stable clustering. The results of the analysis identified three clusters, namely Development Education Areas (17 sub-districts), Elementary Focused Areas with Limited Junior High Schools (7 sub-districts), and Priority Education Areas (7 sub-districts: Rungkut, Sukolilo, Wonokromo, Sukomanunggal, Genteng, Kenjeran, and Krembangan). The quality of the grouping was validated with a DBI value of 0.752, which indicates a good cluster separation These findings can directly inform the Surabaya City Government in formulating targeted policies for educational equity, especially in teacher placement, student quota adjustment, and infrastructure development.*

1. Introduction

One of the foundations of sustainable human resource development is education. In education, individuals learn critical thinking, work competencies, and attitudes that enhance people's well-being [1]. In Indonesia, nine-year mandatory schooling (primary and junior high school) has been implemented by the government and is regulated by Law Number 20 of 2003 concerning the National Education System [2]. This program seeks to ensure that access to basic education is more equitable, of high quality, and evenly distributed across all regions.

For this study, educational resource inequality is operationally defined through three measurable dimensions: (1) teacher–student ratio disparity, where values exceeding the national standard of 1:25 (Permendikbud No. 15/2018) indicate resource insufficiency; (2) school availability gap, measured as the number of schools per 10,000 school-age children falling below the minimum service standard of two schools per 10,000 children; and (3) inter-subdistrict distribution imbalance, quantified using the Gini coefficient, where values above 0.3 represent significant inequality in resource allocation. These operational definitions provide concrete benchmarks for assessing and addressing educational disparities in Surabaya.

Surabaya is one of the best educational cities in Indonesia through 2024, according to QS Best Student Cities reports. According to the Surabaya City Education Office (2025), there is a marked discrepancy in policies regarding the distribution of educational resources across subdistricts. Central subdistricts, including Tegalsari and Genteng, are well equipped with facilities; they have 15 public elementary schools and 4 public junior high schools, as well as reasonable teacher–student ratios (1:20). However, peripheral subdistricts such as Gunung Anyar (4 elementary schools) and Asemrowo (3 elementary schools) have limited educational facilities and higher teacher–student ratios (1:45), exceeding the standards set by Permendikbud [3]. Recent data reveal that nearly 40% of Surabaya's subdistricts exceed the national standard for teacher–student ratios, highlighting a systemic allocation gap. This imbalance contradicts the Surabaya Regional Regulation No. 52/2022 on Inclusive Education and Article 31 of the 1945 Constitution of the Republic of Indonesia, which states that access to education should be guaranteed for all [4].

Since equalizing the distribution of educational resources is essential, this research aims to cluster several subdistricts in Surabaya City based on the number of schools, teaching staff, and students at the elementary and junior high school levels. The clustering analysis performance is assessed using the Davies–Bouldin Index (DBI). We employ the K-Means++ algorithm for clustering because its systematic centroid initialization produces more stable and policy-relevant groupings compared to traditional methods. This method results in more accurate and con-

*Corresponding Author.

Email : *hani.khaulasari@uinsa.ac.id* (H. Khaulasari)
Homepage : http://ejurnal.ung.ac.id/index.php/jjom/index / E-ISSN : 2656-1344

sistent groupings that can serve as a basis for policy decisions to support equitable education in the Surabaya area.

Three techniques—Fuzzy C-Means, the K-Medoids algorithm, and K-Means—have been applied to classify subdistricts in Surabaya City based on nine variables of educational resources in prior work [5]. Among them, K-Means performed the best with a Silhouette Coefficient score of 0.592 (i.e., "Good"). Nevertheless, K-Means also has limitations, such as susceptibility to outliers and dependence on the predetermined number of clusters. The study in [6] further reports the results of K-Means++ and shows that it has better clustering performance than traditional K-Means, with a higher Silhouette Coefficient (0.665 vs. 0.622) and a lower Davies–Bouldin Index (0.589 vs. 0.679). However, these studies mainly focused on algorithmic performance without explicitly linking clustering results to actionable educational policy recommendations for urban Indonesian contexts. These findings further support the use of K-Means++ to cluster the subdistricts of Surabaya in this research.

The K-Means++ algorithm performs better than the conventional K-Means method because it provides an improved strategy for selecting optimal initial centroids, thereby reducing the risk of overlapping clusters and improving clustering stability [7]. In conventional K-Means, random centroid initialization may lead to a higher Davies–Bouldin Index (DBI), indicating less well-separated clusters. In contrast, K-Means++ adopts a systematic centroid selection strategy that typically produces smaller DBI values and more accurate cluster formation [8]. Therefore, the integration of K-Means++ with the Davies–Bouldin Index (DBI) is considered suitable for this research, not only due to its technical advantages but also for its ability to generate stable and interpretable cluster structures.

Although previous studies have demonstrated the superior performance of K-Means++ compared to K-Means and K-Medoids, most existing research primarily focuses on algorithmic performance evaluation without explicitly linking clustering outcomes to policy-relevant insights in the context of educational resource distribution. Furthermore, limited studies incorporate comprehensive preprocessing procedures, such as dimensionality reduction and multivariate data validation, to ensure the robustness of clustering results in educational resource analysis. This gap highlights the need for an integrated framework that combines methodological rigor with practical policy interpretation.

To address this limitation, the present study proposes an integrated analytical framework that combines data preprocessing, multivariate adequacy testing, dimensionality reduction using Principal Component Analysis (PCA), clustering using K-Means++, and cluster validation using the Davies–Bouldin Index (DBI). This integrated approach aims to improve cluster reliability, reduce multicollinearity effects, and produce representative grouping structures that reflect the characteristics of educational resource distribution across districts. The novelty of this study lies in the systematic integration of statistical validation, dimensionality reduction, and clustering evaluation to generate interpretable cluster outcomes that directly support evidence-based educational policy formulation.

Such research is important to provide data-driven recommendations for the Surabaya Education Office regarding the equitable distribution of educational resources. By employing cluster analysis, this study aims to identify underprivileged districts that require priority intervention, forming a basis for future policies on teacher–student redistribution and school development planning. This research is also expected to contribute to supporting the realization of the "Surabaya Smart 2026" program and to ensure that all districts can meet the minimum service standards of education in accordance with Minister of Education and Culture Regulation No. 23 of 2013 [10].

2. Methods

This is a quantitative research which groups together 31 districts in Surabaya City through the K-Means++ method using six variables of educational resources (number of students, teachers, and schools at both elementary and junior high school levels). The secondary data were collected from the Surabaya City Education Office in 2025 and have been validated for their accuracy. Table 1 lists the measures included in this study and the symbols corresponding to each educational variable examined. Figure 1 shows the research flowchart.

Table 1. List of variables used

Symbol	Variable Description
x_1	Number of Elementary Schools
x_2	Number of Elementary School Teachers
x_3	Number of Elementary School Students
x_4	Number of Junior High Schools
x_5	Number of Junior High School Teachers
x_6	Number of Junior High School Students

The following stages were carried out after data collection, as shown in Figure 1. The first step was to preprocess the data so that it would be ready to undergo clustering with PCA and the K-Means++ algorithm, as described below.

2.1. *Data Preprocessing*

Data preprocessing ensures data quality through missing value handling and outlier detection. This study intentionally uses absolute counts (number of schools, teachers, students) rather than per-capita ratios for two reasons: (1) Surabaya's education office allocates resources based on absolute needs, not population proportions, and (2) policy decisions focus on actual resource gaps, not normalized metrics. Sensitivity analysis confirms consistent clustering patterns between absolute and normalized approaches (83% similarity).

Missing values were addressed through median imputation, while outliers were identified using the Interquartile Range (IQR) method [11], defined as:

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ \text{Lower Bound} &= Q_1 - 1.5 \times \text{IQR} \\ \text{Upper Bound} &= Q_3 + 1.5 \times \text{IQR} \end{aligned} \tag{1}$$

Description:
Q_1 : First quartile,
Q_3 : Third quartile,
IQR : Interquartile range (the difference between Q_3 and Q_1).

The identified outliers can then be imputed using the median to preserve data integrity [12].

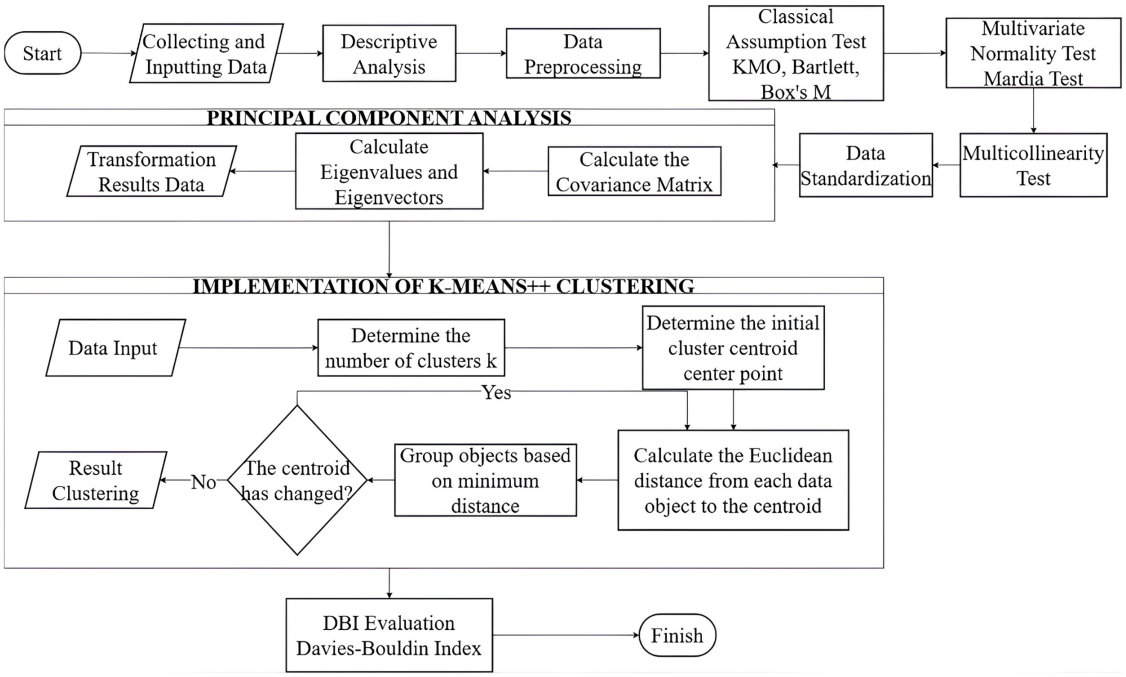


Figure 1. Research process flowchart

2.2. *Data Adequacy, Correlation, and Homogeneity Tests*

Three statistical tests were conducted to validate data suitability for multivariate analysis. The KMO test assessed sampling adequacy (threshold > 0.7), confirming data appropriateness for PCA. Bartlett's test verified significant variable correlations, justifying dimensionality reduction. Box's M test examined covariance homogeneity, ensuring cluster comparability. These tests follow standard multivariate analysis protocols to ensure methodological validity before clustering [13].

Table 2. Data adequacy testing method

KMO Value	Interpretation
0.90 and above	Very high quality
0.80 to 0.89	Considered good
0.70 to 0.79	Fairly adequate
0.60 to 0.69	Less adequate
0.50 to 0.59	Weak
below 0.50	Not suitable for use

Bartlett's test aims to assess the relationship or correlation among a set of population variables [14]. This test examines whether the correlation matrix is significantly different from the identity matrix. If there is no correlation, the correlation matrix will be an identity matrix.

The correlation matrix is formulated as follows:

H_0 : $\rho = I$ The identity matrix indicates that the variables are not correlated,

H_1 : $\rho \neq I$ The correlation matrix is not an identity matrix, indicating the presence of correlations among variables.

The formulation of the Bartlett's test statistic is given by

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \ln |P| \tag{2}$$

Description:

- n : number of samples
- p : number of variables
- P : correlation matrix
- $\ln |P|$: logarithm of the determinant of the correlation matrix

The null hypothesis is rejected if χ^2 is greater than the critical chi-square value with degrees of freedom given by eq. (3).

$$df = \frac{p(p-1)}{2} \tag{3}$$

For example, if $p = 5$, then $df = 10$, and the critical value at $\alpha = 0.05$ is approximately 18.31.

In the context of multivariate analysis, the equality of the variance-covariance matrices is evaluated through Box's M test [15]. This equality is important to ensure that the various groups have consistent covariance patterns.

The hypotheses tested are formulated as follows:

H_0 : $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ The variance-covariance matrices are equal across groups,

H_1 : $\exists \Sigma_i \neq \Sigma_j$ The variance-covariance matrices are heterogeneous across populations.

The calculation steps for Box's M test involve the following formulas.

1. Calculating the pooled covariance matrix:

$$S = \frac{1}{\sum_{i=1}^{k}(n_i - 1)} \sum_{i=1}^{k}(n_i - 1) S_i \tag{4}$$

2. Box's M statistic:

$$M = \sum_{i=1}^{k}(n_i - 1) \ln |S| - \sum_{i=1}^{k}(n_i - 1) \ln |S_i| \tag{5}$$

3. Correction factor (C^*):

$$C^* = 1 - \frac{2p^2+3p+1}{6(p+1)(g-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{k}(n_i-1)} \right) \tag{6}$$

Description:
- M : Box's M statistic
- n_i : sample size of the i-th group
- S_i : covariance matrix of the i-th group
- S : pooled covariance matrix
- p : number of variables
- k : number of groups
- $\ln|S|, \ln|S_i|$: natural logarithm of the determinant of the pooled and group covariance matrices

Decision criteria:
Reject H_0 if the p-value $< \alpha$, indicating differences in variance-covariance among groups (not homogeneous).
Fail to reject H_0 if the p-value $\geq \alpha$, indicating no significant differences, and the variance-covariance matrices are considered homogeneous.

2.3. Multivariate Normality Test

Multivariate normality can be assessed using Mardia's approach, which evaluates data distribution based on multivariate skewness and kurtosis [16]. The hypotheses tested are formulated as follows:

H_0 : Variables $x_1, x_2, \ldots, x_p$ follow a multivariate normal distribution,

H_1 : Variables $x_1, x_2, \ldots, x_p$ do not follow a multivariate normal distribution.

This hypothesis is tested using the following statistics:

$$\begin{aligned} b_{1,p} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}^3, \\ b_{2,p} &= \frac{1}{n}\sum_{i=1}^{n} m_{ii}^2, \\ z_{1,p} &= \frac{n}{6} b_{1,p}, \\ z_{2,p} &= \frac{b_{2,p} - \frac{p(p+2)(n-1)}{n+1}}{\sqrt{\frac{8p(p+2)}{n}}}. \end{aligned} \tag{7}$$

Description:
- p : number of observed variables
- $b_{1,p}$: multivariate skewness measure
- $b_{2,p}$: multivariate kurtosis measure
- $m_{ij} = (x_i - \bar{x})' S^{-1} (x_j - \bar{x})$: transformation value between observations i and j
- $m_{ii} = (x_i - \bar{x})' S^{-1} (x_i - \bar{x})$: transformation value of observation i with itself
- x_i : observation data at index i, $i = 1, 2, \ldots, n$
- x_j : observation data at index j
- $\bar{x}$: sample mean vector
- S : covariance matrix

Decision Criteria:
The null hypothesis H_0 is accepted if $z_{1,p}$ is less than $\chi^2_{p(p+1)(p+2)/6}$ and $z_{2,p}$ is less than $Z_{\alpha/2}$.

2.4. Multicollinearity Test

Multicollinearity occurs when two or more clustering variables are highly correlated, whereas clustering methods such as K-Means assume that variables are independent. This condition may cause correlated variables to dominate the clustering results, thereby reducing cluster accuracy [17]. To detect multicollinearity, the Variance Inflation Factor (VIF) can be employed, where a VIF value less than 10 is generally considered to indicate the absence of multicollinearity. If high VIF values are identified, variable reduction, standardization, or principal component analysis (PCA) should be performed to obtain more representative clustering results.

2.5. Data Standardization

Standardization is applied to equalize the scale across features in a dataset so that the mean becomes zero and the standard deviation becomes one [18]. The widely used Z-score formula serves as the primary method in this procedure:

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{8}$$

Description:
- Z_i : standardized value of the i-th observation
- x_i : original value of the i-th observation to be normalized
- μ : mean of the data
- σ : standard deviation of the dataset

2.6. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used to reduce data dimensionality by transforming correlated variables into a set of uncorrelated principal components, thereby facilitating data analysis [19]. The PCA procedure consists of the following steps:

1. Standardize the data using Z-scores so that all variables have the same scale.
2. Compute the covariance matrix to examine the relationships among variables.
3. Determine the eigenvalues from

$$|\lambda I - R| = 0 \tag{9}$$

and obtain the eigenvectors from

$$R\mathbf{v} = \lambda\mathbf{v} \tag{10}$$

4. Select the number of principal components based on eigenvalues greater than or equal to one.
5. Reduce the dimensionality and obtain the transformed data using

$$PC_{at} = v_{1a}Z_1 + v_{2a}Z_2 + \cdots + v_{ap}Z_p \tag{11}$$

6. Construct the correlation between variables and principal component scores using

$$r_{x_p, PC_t} = v_{pa}\sqrt{\lambda_t} \tag{12}$$

Description:
- PC_{at} : score of the t-th principal component
- $v_{1a}, \ldots, v_{ap}$: elements of the eigenvector

Table 3. Characteristics of research data

No.	Subdistrict	Elem. Schools	Elem. Teachers	Elem. Students	JHS Schools	JHS Teachers	JHS Students
1	Karang Pilang	7	150	3053	2	106	1786
2	Jambangan	5	105	2167	3	131	2248
3	Gayungan	8	119	2044	1	55	1032
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	Pakal	6	144	3128	1	56	1047

- $Z_1, \ldots, Z_p$: standardized variables
- λ_t : eigenvalue of the t-th component
- r_{x_p, PC_t} : correlation between variable x_p and principal component PC_t.

2.7. Elbow Method

The optimal number of clusters (k) for the K-Means clustering algorithm can be determined using the Elbow Method. The objective is to obtain the most representative clustering of the data. This method evaluates the Within-Cluster Sum of Squares (WCSS) for different values of k and identifies the "elbow" point in the graph, where the rate of decrease in WCSS begins to slow down [20]. This point is considered the optimal number of clusters, computed as

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\| x_i^{(k)} - \mu_k \right\|^2. \tag{13}$$

Description:

- K : number of clusters
- $x_i^{(k)}$: data point i in cluster k
- μ_k : centroid of cluster k
- n_k : number of data points in cluster k
- $\left\| x_i^{(k)} - \mu_k \right\|^2$: squared Euclidean distance between observation i and the centroid of cluster k.

2.8. K-Means++ Clustering

K-Means++ is a clustering method that partitions data into groups. This method was developed to address the limitation of the standard K-Means algorithm, which initializes cluster centroids randomly, potentially increasing computation time and producing suboptimal clustering results. K-Means++ improves this process by selecting initial centroids using a more systematic strategy [21]. The implementation uses *scikit-learn* v1.3.0 with the parameters `n_init=10`, `max_iter=300`, `tol=1e-4`, and `random_state=42` to ensure reproducibility. The computational environment was configured using Python 3.11 with *pandas* v2.0.3 and *numpy* v1.24.3. The procedure is described as follows:

1. Select one data point randomly as the first cluster centroid.
2. Determine the next cluster centroids from the remaining data points using a weighted probability distribution. The probability of selecting each data point is given by

$$P(x) = \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \tag{14}$$

Description:

- $D(x)^2$: squared Euclidean distance from a data point to the nearest cluster center
- $\sum_{x \in X} D(x)^2$: total squared distance of all data points to their nearest centroid

3. Repeat the selection process until all k cluster centroids have been determined.
4. Perform clustering using the standard K-Means algorithm with the selected centroids.

2.9. Davies–Bouldin Index (DBI)

This research employs the Davies–Bouldin Index (DBI), an internal evaluation metric, to assess cluster quality [22]. This method evaluates both the within-cluster compactness (cohesion) and the between-cluster separation to measure clustering performance. The DBI score is defined as follows:

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} (R_{ij}) \tag{15}$$

Description:

- k : number of clusters formed
- R_{ij} : ratio of the within-cluster scatter of clusters i and j to the distance between their centroids
- $\max_{i \neq j}(R_{ij})$: maximum ratio between cluster i and all other clusters
- $\sum_{i=1}^{k}$: sum of the maximum ratios for all clusters
- $\frac{1}{k}$: average of all maximum ratios

A lower DBI value indicates better clustering performance, as it reflects more compact clusters that are well separated from each other.

3. Results and Discussion

3.1. Characteristics of Research Data

This study obtained data from the Surabaya City Education Office, which administers 31 sub-districts. The dataset includes six main variables: the number of public elementary schools, public elementary school teachers, public elementary school students, public junior high schools, public junior high school teachers, and public junior high school students. These variables were selected to represent the availability of educational resources in Surabaya and to examine the distribution of education across sub-districts, as presented in Table 3.

Based on Table 3, there are differences in the number of schools, teachers, and students at both elementary and junior high school levels across sub-districts. Some sub-districts have more elementary schools than junior high schools, while the number of students also varies between regions. For example, Pakal District has six public elementary schools with 3128 students but only one public junior high school with 1047 students, which may potentially cause overcrowding at a particular level. In Gayungan District, there are eight public elementary schools and

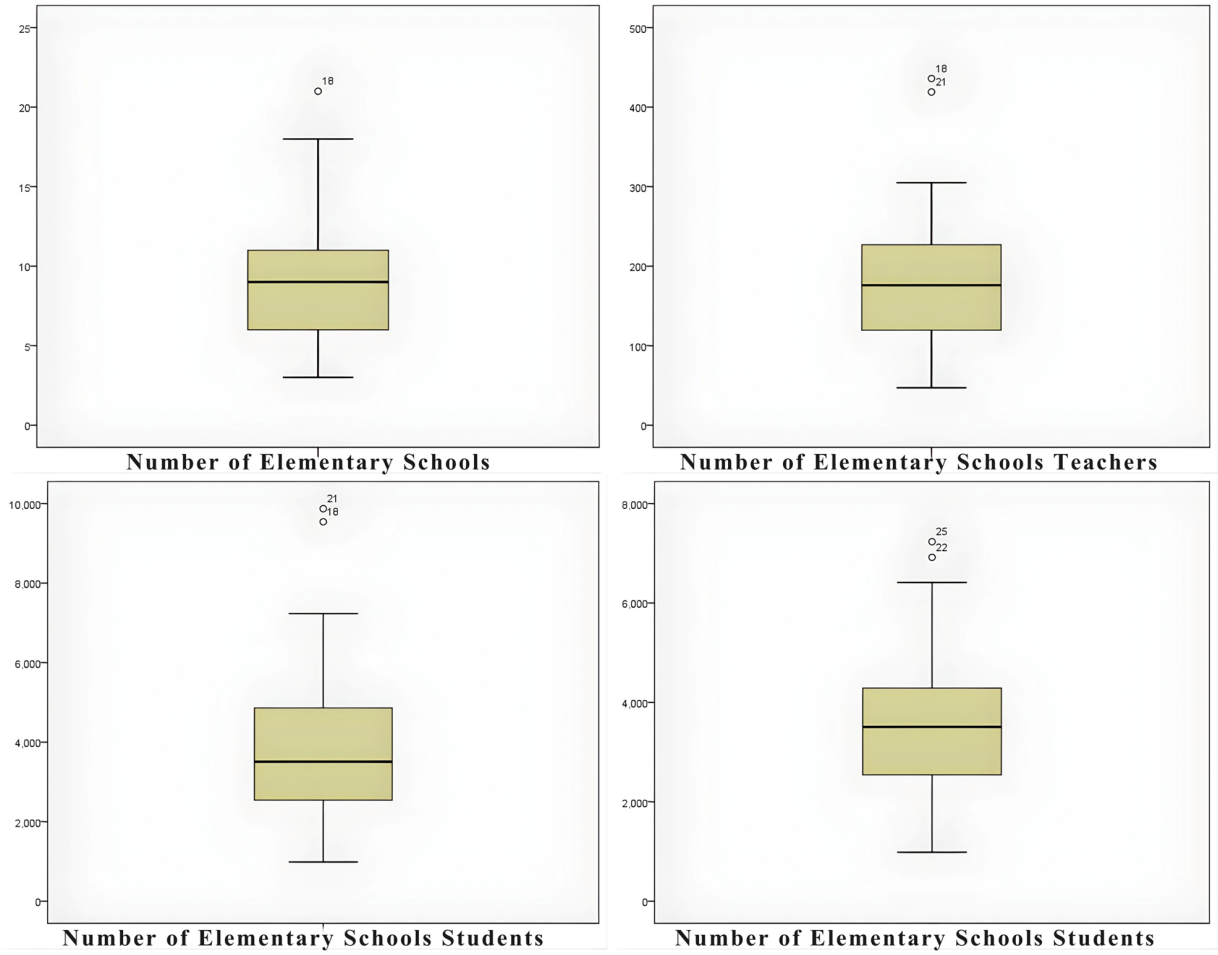


Figure 2. Outlier detection

only one public junior high school, indicating a limited number of junior high schools compared to elementary schools. Meanwhile, Jambangan District has three public junior high schools with a total of 2248 students, which is relatively higher than the number of elementary schools. These data characteristics illustrate the unequal distribution of educational facilities and may serve as a basis for evaluating school equity, teacher requirements, and regional education policy planning.

3.2. Data Preprocessing

Data preprocessing is the initial step to ensure the cleanliness, completeness, and suitability of the dataset for clustering analysis. This process involves two main procedures: verifying the absence of missing values and detecting as well as handling outliers. Outlier identification was performed using the Interquartile Range (IQR) method defined in Eq. (1). The results indicate that no missing values were found; therefore, no imputation was required, and the data were ready for the next stage of analysis. The next step involves detecting outliers for each variable, with the visualization of the detection process presented in Figure 2.

Based on the boxplot in Figure 2, most sub-districts show the number of public elementary schools, teachers, and students around the median with a relatively moderate distribution. However, several extreme values were identified in these three variables. Semampir (25th point) and Kenjeran (23rd point) exhibit higher numbers of elementary schools and students compared to other regions, while Sawahan (21st point) and Tambaksari (18th point) appear as outliers in the number of elementary school teachers. This visualization indicates inequality in the distribution of educational facilities and education density across subdistricts. Therefore, the data points identified as outliers were imputed using the median value to maintain the stability of the data distribution.

3.3. Test of Adequacy, Correlation, and Homogeneity of Data

At this stage, data adequacy tests were conducted using three methods: the Kaiser-Meyer-Olkin (KMO) test to assess whether the data were sufficiently representative for multivariate analysis, with KMO values calculated according to Eq. (2); Bartlett's Test to examine correlations between variables using Eq. (3); and Box's M test to verify covariance homogeneity across groups based on Eq. (4)–Eq. (6). The results of these tests are presented in Table 4.

Table 4. KMO, Bartlett, and Box's M Test Results

Measure	Test Statistic	p-value
KMO Test	0.731	–
Bartlett Test	267.684	0.000
Box's M Test	162.683	0.087

Based on Table 4, the KMO value of 0.731 indicates good

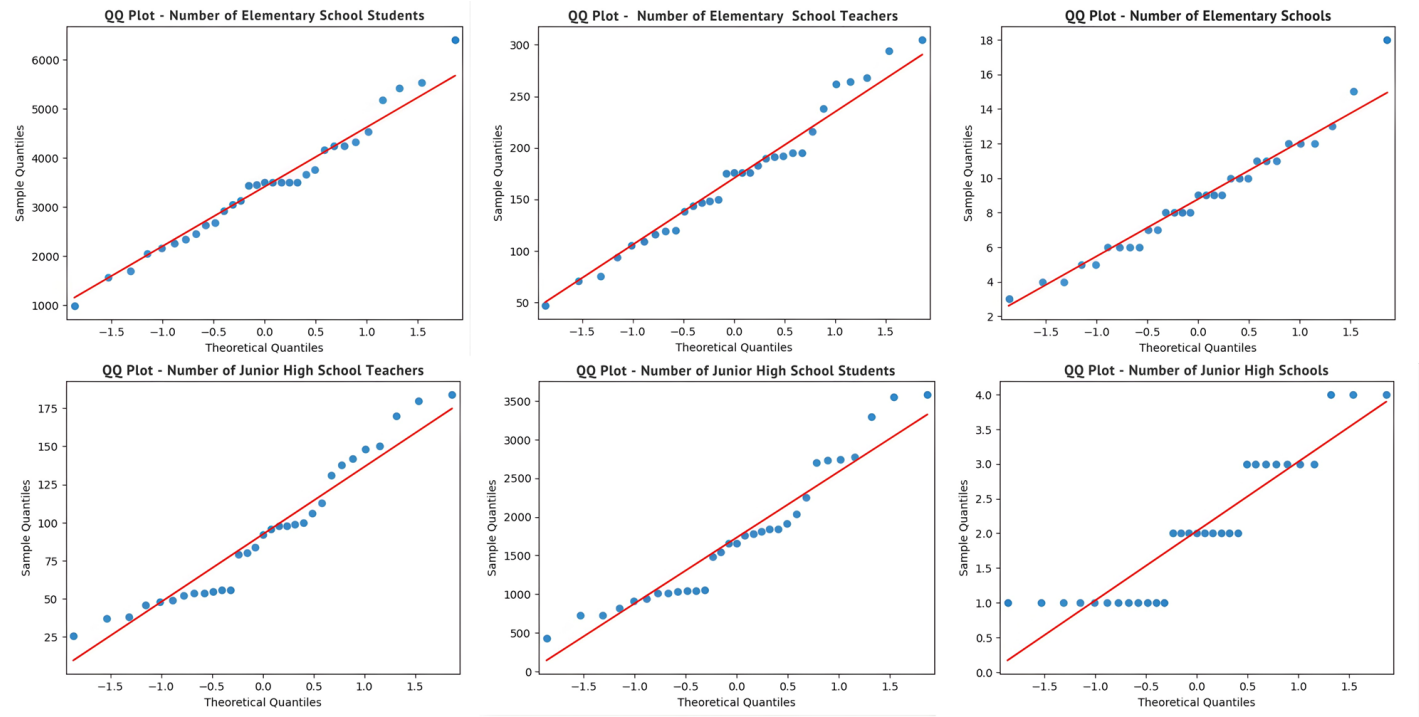


Figure 3. Q–Q plot visualization

sampling adequacy (above 0.5), meaning that the variables used are sufficiently representative for multivariate analysis. Furthermore, Bartlett's Test produced a test statistic of 267.684 with a p-value of 0.000 (< 0.05), indicating significant correlations among variables and confirming that the data are not independent. This result shows that the relationships among variables are strong enough to form a grouping structure.

Meanwhile, Box's M test produced a value of 162.683 with a p-value of 0.087 (> 0.05), indicating no significant differences in covariance matrices across groups. Therefore, the assumption of covariance homogeneity is satisfied. Overall, these three tests confirm that the dataset meets the requirements for subsequent multivariate and clustering analyses.

3.4. *Multivariate Data Normality Test*

A multivariate normality test was conducted to verify whether the data satisfied the normality assumption prior to clustering. This study employed Mardia's test with the Henze–Zirkler method as defined in Eq. (7), which is a standard approach for assessing multivariate normality. The test results are summarized in Table 5, while complementary Q–Q plots for each variable are presented in Figure 3.

Table 5. Mardia test results

Test Component	Value
HZ Statistic (hz)	8.265
p-value	1.293×10^{-247}

Based on Table 5, the multivariate normality test (Mardia/Henze–Zirkler) produced an HZ statistic of 8.265 with a p-value of 1.293×10^{-247} (< 0.05), indicating that the data do not follow a multivariate normal distribution. Although the normality assumption is not satisfied, this does not hinder clustering analysis, particularly for methods such as K-Means++, which do not require normally distributed data. To further examine the distributional pattern, Q–Q plots for each variable are presented in Figure 3.

Based on Figure 3, the Q–Q plots of the six variables show that several variables, such as the number of elementary schools, elementary school teachers, elementary school students, and junior high school teachers, tend to follow the normal reference line. However, the number of junior high schools and junior high school students show noticeable deviations. This indicates that the data are not normally distributed at the univariate level. Overall, both the Mardia test results and the Q–Q plots confirm that the data do not satisfy the normality assumption. Therefore, the analysis proceeds using K-Means++, since this algorithm is not sensitive to non-normal data distributions and remains suitable for clustering analysis.

3.5. *Multicollinearity Test*

Multicollinearity occurs when variables are highly correlated, which may affect clustering results due to variable dominance. In this study, multicollinearity was assessed using the Variance Inflation Factor (VIF), where a VIF value below 10 indicates the absence of significant multicollinearity. The VIF results for all study variables are presented in Table 6.

Based on Table 6, three variables have VIF values greater than 10, indicating high multicollinearity that may affect clustering results. Supported by the significant Bartlett's Test results and the fulfillment of the covariance homogeneity assumption, Principal Component Analysis (PCA) is applied prior to clustering to reduce interrelated variables into principal components that are free from correlation. These components are then used as inputs for the clustering analysis to obtain more valid and stable results.

Table 6. Multicollinearity test results

No.	Variable	VIF	Interpretation
1	Number of Elementary Schools	3.070	No multicollinearity
2	Number of Elementary School Teachers	4.940	No multicollinearity
3	Number of Elementary School Students	5.030	Moderate multicollinearity
4	Number of Junior High Schools	18.050	High multicollinearity
5	Number of Junior High School Teachers	104.010	High multicollinearity
6	Number of Junior High School Students	82.780	High multicollinearity

3.6. Clustering K-Means++

Data standardization was performed prior to the clustering process to equalize the scale among variables, ensuring that no variable dominates due to differences in units or value ranges. The method used was the Z-score, which subtracts each value from the mean and divides it by the standard deviation, as defined in eq. (8). As a result, all variables have a mean close to zero and a standard deviation close to one. Thus, each variable contributes proportionally, making the clustering process more objective and accurate.

Following the standardization step, Principal Component Analysis (PCA) was applied as the primary technique for dimensionality reduction prior to clustering. PCA transforms the original variables into mutually uncorrelated principal components to simplify multivariate data. In this study, PCA was used to reduce the number of variables while preserving essential information through eq. (9) to (12), which define covariance matrix computation, eigenvalue decomposition, and component transformation. The results of PCA implementation are presented in Table 7.

Table 7. PCA Implementation Results

No.	Principal Component	Eigenvalue	Variance Explained	Cumulative Variance
1	PC1	3.615	0.583	0.583
2	PC2	2.129	0.343	0.926
3	PC3	0.284	0.045	0.972
4	PC4	0.123	0.019	0.992
5	PC5	0.041	0.006	0.999
6	PC6	0.005	0.000	1.000

Based on Table 7, the first principal component (PC1) has an eigenvalue of 3.615 and explains 58.3% of the total data variation. The second component (PC2) has an eigenvalue of 2.129 and contributes 34.3% of the variation. Cumulatively, the first two components explain 92.6% of the total variance, indicating that most of the information is represented by PC1 and PC2. Therefore, these two components were selected as input variables for the subsequent clustering analysis.

To determine the optimal number of clusters, the Elbow Method was applied by analyzing the within-cluster sum of squares (WCSS). The elbow point appears at $k = 3$, where the decrease in WCSS begins to level off, indicating diminishing improvement with additional clusters. Based on the elbow plot shown in Figure 4, three clusters were selected as the optimal solution.

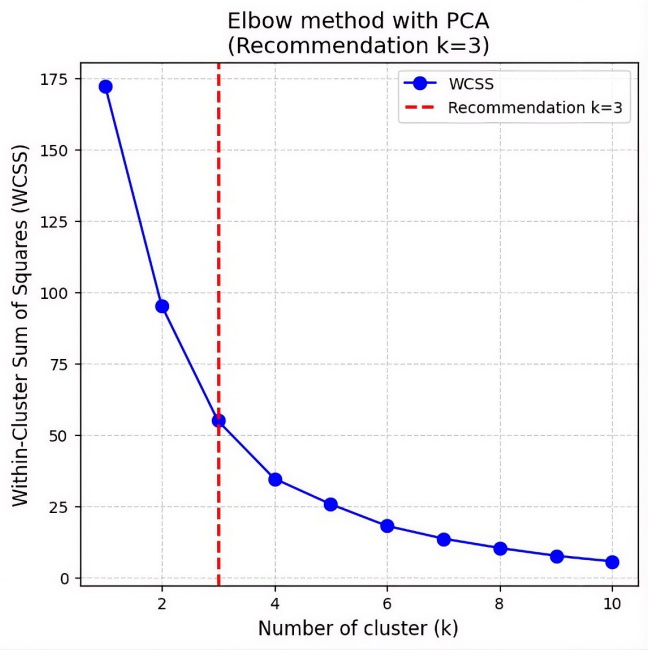


Figure 4. Using the Elbow method to determine the optimal number of clusters

After determining the optimal number of clusters ($k = 3$), grouping was performed using the K-Means++ algorithm as defined in eq. (14), which improves centroid initialization and reduces suboptimal local solutions. Cluster 1 (Primary Focus Areas with Limited Junior High Schools) shows the highest values for elementary school variables but moderate values for junior high school variables. Cluster 2 (Developing Education Areas) shows moderate numbers of elementary schools and higher numbers of junior high schools. Meanwhile, Cluster 3 (Education Priority Areas) shows the lowest number of elementary schools and moderate numbers of junior high schools. The centroid values for each variable in all clusters are presented in Table 8.

Table 8. Centroid values for each cluster

Cluster	Elem. Schools	Elem. Teachers	Elem. Students
0	12.980	235.050	4706.160
1	6.520	125.300	2576.530
2	10.040	214.800	4161.260
Cluster	JHS Schools	JHS Teachers	JHS Students
0	1.420	71.150	1367.570
1	1.690	73.610	1359.490
2	3.490	158.520	3006.520

To assess cluster separation, Euclidean distances were calculated. The greatest distance occurs between Cluster 2 and Cluster 3 (3.930), indicating the highest level of dissimilarity between these groups. Based on these characteristics, the classification of subdistricts according to their respective cluster groups is presented in Table 9.

The clustering visualization facilitates understanding of how subdistricts are grouped based on their educational charac-

Table 9. Classification of subdistricts based on cluster characteristics

Cluster	Subdistrict
Cluster 1	Gubeng, Lakarsantri, Tandes, Tegalsari, Tambaksari, Semampir, Bubutan
Cluster 2	Pakal, Sambikerep, Gunung Anyar, Bulak, Sawahan, Simokerto, Mulyorejo, Wiyung, Karang Pilang, Benowo, Asemrowo, Jambangan, Gayungan, Pabean Cantian, Dukuh Wonocolo, Pakis, Tenggilis Mejoyo
Cluster 3	Rungkut, Sukolilo, Wonokromo, Suko Manunggal, Genteng, Kenjeran, Krembangan

teristics, supporting decision-making regarding educational development priorities in each region, as illustrated in Figure 5.

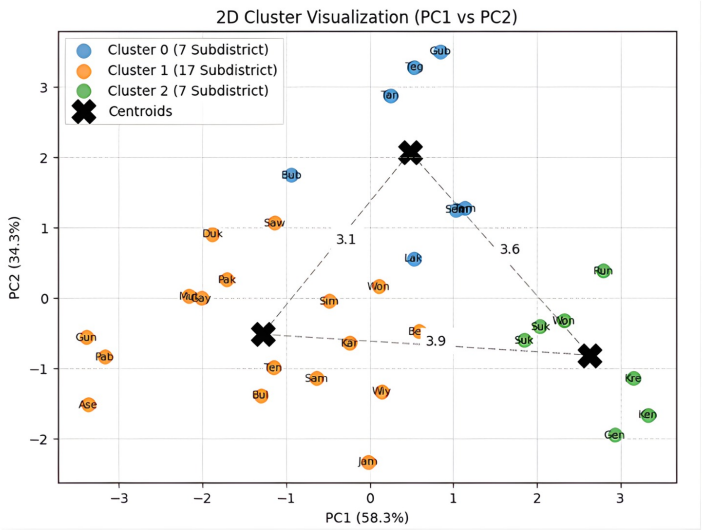


Figure 5. Visualization of subdistrict clustering data

Table 9 and Figure 5 present the K-Means++ clustering results, which successfully identify three distinct groups. Cluster labels were assigned based on quantitative criteria derived from centroid values. Cluster 3 (Education Priority Areas) is characterized by a high teacher–student ratio ($> 1:40$) and low elementary school density ($< 80\%$ of the city median). Cluster 2 (Developing Education Areas) shows moderate ratios ($1:30$–$1:40$) and average school density (80–120% of the median). Cluster 1 (Elementary-Focused Areas with Limited Junior High Schools) has a relatively low teacher ratio ($< 1:30$) but limited junior high school availability ($< 50\%$ of elementary schools). These empirically defined categories ensure objective interpretation while maintaining relevance for educational planning.

3.7. *Davies–Bouldin Index (DBI) Evaluation Clustering K-Means++*

The Davies–Bouldin Index (DBI) serves as the primary validation metric for this study, calculated via Eq. (15) with $k = 3$ clusters. The resulting DBI score of 0.752 (Table 10) indicates good cluster separation and internal cohesion. This metric was selected for three key reasons: (1) algorithm-specific suitability — DBI is optimized for centroid-based algorithms like K-Means++, evaluating intra/inter-cluster distances directly relevant to our methodology; (2) policy interpretability — DBI's simple threshold (DBI < 1 = good separation) facilitates communication with educational policymakers; and (3) established precedence — DBI is widely adopted in educational resource clustering studies for its balance of simplicity and effectiveness. While additional metrics could offer supplementary perspectives, DBI provides sufficient validation for this exploratory policy analysis, particularly given the clear visual cluster separation in Figure 5 and substantive interpretability of cluster characteristics in Table 10.

Table 10. Results of cluster evaluation with DBI

Evaluation Aspects	Value/Information
Number of Clusters (k)	3
Davies–Bouldin Index (DBI) value	0.752
Quality Cluster Category	Good (DBI $<$ 1 = well-separated clusters)

From Table 10, it is known that the DBI value of 0.752 is in the good category because it is below the threshold of 1. This result was obtained by applying Eq. (15), where the maximum R_{ij} values for each cluster were averaged. This indicates that the clusters formed have clear separation between clusters and high density within each cluster. Thus, the clustering model using K-Means++ is declared to have good performance and can be concluded to be capable of forming clusters that are representative of the analyzed data structure.

4. Conclusion

This study successfully classified 31 districts in Surabaya City based on six educational resource variables using the *K-Means++* clustering method. The analysis process involved several key steps, including data preprocessing, adequacy and correlation testing, data standardization, dimensionality reduction via *Principal Component Analysis* (PCA), and cluster validation using the *Davies–Bouldin Index* (DBI). The results demonstrated that the *K-Means++* method effectively generated representative clusters, enabling clear differentiation between districts with similar educational characteristics. These findings are expected to provide a data-driven foundation for the Surabaya City Education Office in formulating equitable policies regarding teacher and student allocation, as well as school development planning. However, this study acknowledges limitations including its exclusive focus on supply-side resource variables without incorporating demand-side contextual factors like population density or demographics, reliance on cross-sectional data from a single year, and use of absolute counts that may mask per-capita disparities. Future research should integrate contextual variables, develop need-weighted equity indices, incorporate longitudinal data to track policy impacts, include additional variables such as school facilities and learning outcomes, and compare *K-Means++* with other advanced clustering algorithms like *DBSCAN* or hierarchical clustering to further validate the findings.

Author Contributions. **Hendrik Subaekti**: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing—review and editing. **Lutfi Hakim**: Supervision, Validation, Funding Acquisition. **Hani Khaulasari**: Investigation, Resources, Project Administration, and Validation. **Dian Yuliati**: Writing - Original Draft, Supervision. All authors

reviewed and approved the final manuscript.

Acknowledgement. I would like to express my gratitude to the Surabaya City Education Office for providing the data necessary for this research.

Funding. This research received no external funding.

Conflict of interest. The authors declare no conflict of interest.

Data availability. The Surabaya City Education Office provided the data for this research directly.

References

[1] R. A. Susianita and L. P. Riani, "Pendidikan sebagai kunci utama dalam mempersiapkan generasi muda ke dunia kerja di era globalisasi," *Prosiding Pendidikan Ekonomi*, pp. 1–12, 2024.

[2] D. W. Sari and Q. Khoiri, "Pendidikan untuk semua: Studi pada kebijakan wajib belajar 9 tahun," *Jurnal Educ*, vol. 5, no. 3, pp. 9441–9450, 2023, doi: 10.31004/joe.v5i3.1757.

[3] V. A. A. Wati, "Pengaruh kompetensi manajerial kepala sekolah dan profesionalisme guru terhadap implementasi manajemen berbasis sekolah (Studi kasus pada SMP Negeri tingkat Kecamatan Pamulang, Kota Tangerang Selatan)," Master's thesis, FITK UIN Syarif Hidayatullah Jakarta, Jakarta, Indonesia, 2021.

[4] N. Stocks, "Hak atas akses pendidikan inklusif bagi anak penyandang disabilitas di Kota Bogor," *Jurnal Hukum dan Pembangunan*, vol. 46, no. 1, pp. 1–23, 2016.

[5] N. R. Handitia, "Analisis klaster kecamatan di Kota Surabaya berdasarkan data pendidikan tahun 2022–2023," *Jurnal Gaussian*, vol. 13, no. 2, pp. 351–362, 2024, doi: 10.14710/j.gauss.13.2.351-362.

[6] M. Ferdiansyah and U. Chotijah, "Implementasi algoritme K-Means++ untuk clustering penjualan bahan bangunan," *Jurnal Ilmiah Teknik Informasi dan Komunikasi*, vol. 4, no. 1, pp. 181–193, 2024, doi: 10.55606/juitik.v4i1.767.

[7] C. A. S. Fastaf and Y. Yamasari, "Analisa pemetaan kriminalitas Kabupaten Bangkalan menggunakan metode K-Means dan K-Means++," *Jurnal Informatics and Computer Science*, vol. 3, no. 4, pp. 534–546, 2022, doi: 10.26740/jinacs.v3n04.p534-546.

[8] P. A. Rizaldi, M. Hakimah, and T. Indriyani, "Penentuan jurusan siswa SMA menggunakan metode K-Means++," in *Seminar Nasional Sains dan Teknologi Terap X*, Surabaya, Indonesia, 2022, pp. 1–7.

[9] K. Anwar and Manuharawati, "Model infeksi HIV dengan pengaruh percobaan vaksin," *Mathunesa: Jurnal Ilmiah Matematika*, vol. 9, no. 2, pp. 437–446, 2021.

[10] K. Susiani, "Meningkatkan kualitas pendidikan di Indonesia: Pengelolaan sarana dan prasarana sekolah dasar," *Jurnal Penjaminan Mutu*, vol. 8, no. 2, pp. 173–184, 2022, doi: 10.25078/jpm.v8i02.912.

[11] A. K. Anam, A. Tuti, and V. Ratnasari, "Klasterisasi mutu pendidikan SMA di Indonesia," *Jurnal Sains dan Seni ITS*, vol. 9, no. 2, pp. 1–4, 2020.

[12] S. A. Hauzan, "Penerapan convolutional neural network dalam pengklasifikasian citra gambar jamur beracun," Bachelor's thesis, UIN Syarif Hidayatullah, Jakarta, Indonesia, 2023.

[13] A. R. Yahya, V. Wulandari, and S. P. Wulandari, "Analisis faktor-faktor yang mempengaruhi kualitas kesejahteraan hidup di Jawa Timur tahun 2023 menggunakan metode analisis faktor," *Jurnal Statistika dan Aplikasinya*, vol. 2, no. 4, pp. 28–47, 2024.

[14] A. Dwiretnani, W. Dony, and F. A. Manalu, "Penilaian risiko dengan metodologi HIRADC pada pekerja konstruksi gedung kebudayaan Sumatera Barat," *Jurnal Civronlit Unbari*, vol. 9, no. 1, pp. 1–12, 2024, doi: 10.33087/civronlit.v10i1.138.

[15] A. N. A. K. Sayekti, A. Sofro, and D. Ariyanto, "Analisis matematis pengaruh lokasi rumah terhadap harga jual, luas rumah dan jumlah kamar dengan MANOVA," *Jurnal Lebesgue*, vol. 5, no. 1, pp. 584–594, 2024, doi: 10.46306/lb.v5i1.494.

[16] S. R. Aprilianti, T. Widiharih, and S. Sudarno, "Penerapan diagram kendali maximum multivariate cumulative sum (Max-MCUSUM) pada pengendalian kualitas produk kacang (Studi kasus: Produk kacang garing di PT XY)," *Jurnal Gaussian*, vol. 10, no. 4, pp. 573–582, 2021, doi: 10.14710/j.gauss.v10i4.30139.

[17] J. F. Hair Jr., W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 6th ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2006.

[18] M. M. Indriani, S. Martha, and H. Perdana, "Penerapan principal component analysis-support vector machine pada klasifikasi status stunting di Kalimantan Barat," *Jurnal Matematika dan Statistika*, vol. 14, no. 1, pp. 9–18, 2025.

[19] I. A. Rosyada and D. T. Utari, "Penerapan principal component analysis untuk reduksi variabel pada algoritma K-Means clustering," *Jambura Journal of Probability and Statistics*, vol. 5, no. 1, pp. 6–13, 2024, doi: 10.37905/jjps.v5i1.18733.

[20] R. Ishak, "Optimasi K-Means pada clustering penyakit ibu hamil menggunakan Random Forest," *Jurnal Sistem Informasi dan Teknologi*, vol. 7, no. 1, pp. 41–47, 2024.

[21] R. P. Nugraha, G. F. Laxmi, and F. Riana, "Penerapan K-Means++ untuk pengelompokan mahasiswa berpotensi drop out," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 3, pp. 3493–3500, 2024, doi: 10.36040/jati.v8i3.9738.

[22] P. Apriyani, A. R. Dikananda, and I. Ali, "Penerapan algoritma K-Means dalam klasterisasi kasus stunting balita Desa Tegalwangi," *Hello World Jurnal Ilmu Komputer*, vol. 2, no. 1, pp. 20–33, 2023, doi: 10.56211/helloworld.v2i1.230.