SAGA: Journal of Technology and Information Systems
Vol 4, Issue 4, November 2026, Page 700-709
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i4.837

Identifying Student Learning Performance Patterns Through K-Means Clustering Based on Multi-Assessment Scores

Roberto Kaban[1*], Sela Sari[2]

[1]Department of Informatics Engineering, Institut Teknologi dan Bisnis Indonesia, Deli Serdang, Indonesia
[2]UPT SPF SD Negeri 104222 Sei Glugur, Deli Serdang, Indonesia

Email: roberto.kaban@yahoo.com[1]*, selasari99@guru.sd.belajar.id[2]
Orcid: https://orcid.org/0000-0002-9785-8382[1]*

Abstract. Teachers routinely collect several assessment scores for each pupil, yet these are usually condensed into a single grade that hides differences in the level and consistency of performance. This study applied K-means clustering to identify learning performance patterns among 36 fourth-grade pupils in the Mekarsi (Melayu, Karo, and Simalungun) subject in the second semester of the 2025–2026 academic year. Twelve scores were consolidated into four standardized features: the means of seven learning-objective and three chapter-level assessments and the test and non-test components of the end-of-semester assessment. Solutions with two to eight clusters were compared using the sum of squared errors, silhouette coefficient, Davies–Bouldin index, and Calinski–Harabasz index. Three clusters were retained for interpretability, although two clusters had the highest silhouette value (0.560). The solution comprised high-performing (n = 7; composite mean 85.68), moderate (n = 14; 81.11), and developing pupils (n = 15; 78.81), with an average silhouette of 0.321, indicating weak structure. The high-performing cluster scored 3.14 points higher on the last than on the first learning-objective assessment, compared with 0.73 points for the developing cluster, a descriptive difference rather than evidence of learning gain. Multi-assessment clustering can reveal patterns that a class average conceals, provided that the labels guide follow-up rather than serve as fixed judgments of children.

Keywords: Educational data mining, K-means clustering, Multi-assessment, Primary education, Student performance

1 Introduction

Assessment is the main mechanism through which teachers monitor what pupils have learned and decide what to teach next, and classroom evidence that is used to adapt instruction is one of the most powerful levers for improving learning [1]. Under the Indonesian Merdeka Curriculum, assessment is organized into diagnostic, formative, and summative forms, and teachers are encouraged to use the evidence gathered during learning, not only end-of-term results, to improve instruction [2]. Research on teachers' perceptions of these assessment changes describes all three forms as part of everyday classroom practice [3]. In practice, a single pupil accumulates scores from several instruments within one semester, for example, assessments tied to individual learning objectives, chapter-level summative assessments, and an end-of-semester examination. These records are routinely collected, but they are usually condensed into a final grade or a class average, which hides how pupils differ in the level and consistency of their performance across instruments.

Educational data mining (EDM) offers one way to recover this hidden structure by applying data mining techniques to data generated in educational settings [4]–[6]. Clustering, an unsupervised technique, groups learners with similar score profiles without predefined labels, and K-means [7], [8] is the algorithm most often chosen because it is simple, fast, and easy to explain to teachers. K-means has been used to segment learners by performance in a national education portal [9], to distinguish low, average, and high performers in higher education [10], to follow how students migrate between clusters over time [11], to compare academic achievement before and during the COVID-19 pandemic [12], and to arrange university students' scores into three layers [13]. The three-group solutions reported in [10] and [13] are noteworthy because they suggest that a coarse division into high, middle, and lower performers is a recurring and practically meaningful way of describing a cohort. Table 1 summarizes the settings and purposes of these studies and shows how the present study differs from them.

Table 1. Summary of related K-means studies of student performance and the position of the present study.

Study	Educational setting	Purpose of clustering	Method
Alalawi *et al.* [9]	National education portal	Segment learners by performance	K-means
Talib *et al.* [10]	Higher education	Distinguish low, average, and high performers	K-means and SVM
Tuyishimire *et al.* [11]	Tertiary education	Follow students' migration between	K-means

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 4, November 2026, Page 700-709
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i4.837

Study	Educational setting	Purpose of clustering	Method
		clusters over time	
Ahmad *et al.* [12]	Higher education	Compare achievement before and during COVID-19	K-means
Abdelmagid and Qahmash [13]	University	Arrange students' scores into three layers	Orange data mining
Irmayanti *et al.* [14]	Elementary school	Profile academic achievement	K-means; elbow and silhouette
This study	Primary school (Grade IV)	Profile multi-assessment performance and its instructional use	K-means; four validity indices and stability checks

Two gaps motivate the present study. The first concerns interpretation. A recent systematic review found that the elbow and silhouette methods are widely used to choose the number of clusters but are often applied procedurally, and that clusters can be statistically valid yet pedagogically uninformative when they are not connected to educational meaning [15]. The second concerns the setting. Most of the studies cited above deal with higher education [10]–[13], whereas primary teachers must interpret the score patterns of minors whose report-card scores are strongly context-dependent and whose labels can shape expectations, a caution raised explicitly in a study of elementary achievement profiles built on K-means with elbow and silhouette evaluation [14]. In addition, the choice of the number of clusters is rarely obvious: the elbow and silhouette approaches are popular, yet both require repeated runs of K-means and have known weaknesses [16], and comparisons of validity indices show that no single index performs best in every situation [17].

This study therefore applies K-means clustering to the multi-assessment scores of one fourth-grade class in the Mekarsi subject. Its objectives are (1) to construct a compact set of features from learning-objective, chapter-level, and end-of-semester assessments; (2) to determine a defensible number of clusters using several validity indices instead of a single criterion; and (3) to describe the resulting performance profiles and discuss their instructional implications and limits. The contribution is not a new algorithm but a transparent, reproducible example of how routine classroom records can be turned into interpretable learner profiles, together with an explicit account of how far those profiles can be trusted.

2 Methodology

The research followed six stages, from data collection to profiling and interpretation, as summarized in Figure 1. Each stage is described in the following subsections.

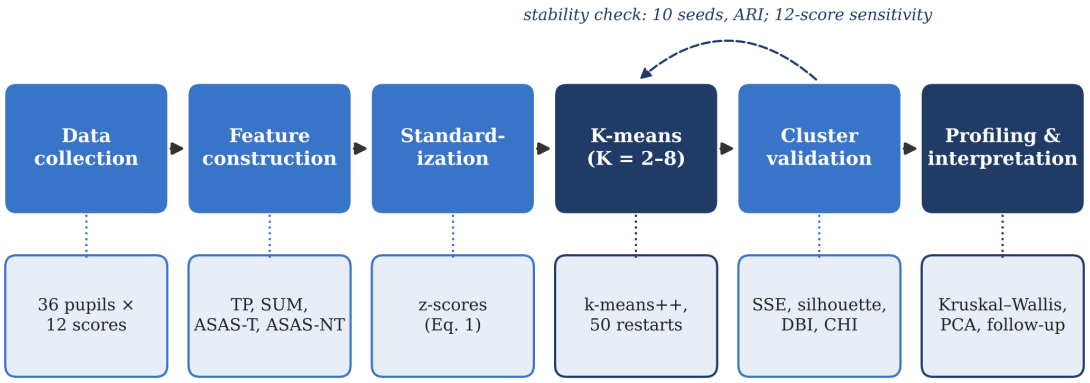

Figure 1. Research workflow of the study.

2.1 Research Design and Data

The study used a quantitative, exploratory, cross-sectional design based on secondary data. The dataset is the official score recapitulation (*Rekapitulasi Nilai*) of the Grade IV Mekarsi (Melayu, Karo, and Simalungun) subject for Semester 2 of the 2025–2026 academic year, obtained from SD Negeri No. 104222 Sei Glugur, and covers all 36 pupils in the class. No pupil was excluded, and no missing values were found. Each record contains 12 assessment scores: seven learning-objective assessments (TP1–TP7), three chapter-level summative assessments (Chapters 1–3), and two components of the end-of-semester summative assessment (ASAS),

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 4, November 2026, Page 700-709
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i4.837

namely a test component (ASAS-T) and a non-test component (ASAS-NT). The three instrument types (TP, chapter-level, and ASAS) were not equated for difficulty or scoring, so score differences between instruments are treated as descriptive throughout. Observed scores ranged from 77 to 94. Before analysis, pupil names were replaced with codes S01–S36 that follow the order of the original register.

2.2 Feature Construction and Preprocessing

Clustering twelve strongly related scores for only 36 pupils would give the instrument types with the most assessments a disproportionate weight, so the scores were aggregated into four features that mirror the structure of the assessment system (Table 2). Aggregation dampens the noise of any single assessment and keeps the ratio of observations to variables reasonable.

Table 2. Definition of the clustering features.

Feature	Source assessments	No. of scores	Description
TP	TP1–TP7	7	Mean of the learning-objective assessments
SUM	Chapters 1–3	3	Mean of the chapter-level summative assessments
ASAS-T	ASAS test	1	Test component of the end-of-semester summative assessment
ASAS-NT	ASAS non-test	1	Non-test component of the end-of-semester summative assessment

The four features were then standardized to z-scores using Equation (1) so that no feature dominates the Euclidean distance because of its scale or variance. Because the scores occupy a narrow band, z-scoring stretches small raw differences (a few points) into full standardized units, so the clusters separate pupils on small absolute score differences. It also gives the four features equal weight in the distance calculation even though they are strongly correlated (Section 3.1), so the shared overall-performance dimension dominates the partition (Section 3.3).

$$z_{ij} = (x_{ij} - \bar{x}_j) / s_j \quad (1)$$

Here, x_{ij} is the value of feature j for pupil i, and $\bar{x}_j$ and s_j are the mean and standard deviation of that feature. The overall composite reported in Tables 4 and 7 is the unweighted mean of the four unstandardized features; it is used only to describe and order the clusters and was not an input to the clustering. A sensitivity analysis that uses all twelve scores is reported in Section 3.2. No rank-based or alternative normalization was tested (see Section 3.5).

2.3 K-Means Clustering

K-means partitions the pupils into K clusters by minimizing the within-cluster sum of squared errors (SSE) given in Equation (2), where C_k is the k-th cluster and μ_k is its centroid [7], [18]. The algorithm was run in Python 3.11 with scikit-learn 1.8.0 [19], using k-means++ initialization [20], 50 random restarts, and a fixed random seed of 42; the solution with the lowest SSE across restarts was retained.

$$SSE = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \quad (2)$$

2.4 Cluster Validation and Stability

Because K must be fixed in advance, values from 2 to 8 were evaluated with four criteria: the SSE, inspected for an elbow; the mean silhouette coefficient [21], where the silhouette of pupil i is defined in Equation (3), $a(i)$ being the mean distance to the other members of the pupil's own cluster and $b(i)$ the smallest mean distance to the members of any other cluster; the Davies–Bouldin index [22] (lower is better); and the Calinski–Harabasz index [23] (higher is better). Since each criterion has limitations [16] and none is superior in all situations [17], the final choice of K combined the indices with the interpretability of the resulting groups, consistent with the recommendation to use validation methods complementarily [15].

$$s(i) = (b(i) - a(i)) / max\{a(i), b(i)\} \quad (3)$$

Stability was examined in three ways: by repeating the final analysis with ten different random seeds (50 restarts each), by repeating it with 100 single-start runs, and by comparing partitions with the adjusted Rand index (ARI) [24]. Robustness to the choice of features was examined by clustering the full twelve-score matrix. Differences among clusters in each feature were summarized with the Kruskal–Wallis test [25], and the clusters were visualized on the first two principal components of the standardized features [26]. Because the clusters were formed from the same features, these tests are descriptive and are not confirmatory evidence of group

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 4, November 2026, Page 700-709
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i4.837

differences. No additional data were collected from pupils, and cluster labels were treated as descriptive summaries of one semester rather than as judgments of ability. Table 3 summarizes the analysis configuration.

Table 3. Configuration of the clustering analysis.

Component	Setting
Software	Python 3.11 with scikit-learn 1.8.0 [19]
Input	36 pupils × 4 standardized features (z-scores)
Algorithm and initialization	K-means (Lloyd) with k-means++ seeding [20]
Restarts and seed	50 random restarts; random seed 42; lowest SSE retained
Candidate number of clusters	K = 2–8
Validity criteria	SSE (elbow), silhouette, Davies–Bouldin, and Calinski–Harabasz indices
Stability checks	10 seeds × 50 restarts; 100 single-start runs; adjusted Rand index
Sensitivity analysis	Clustering of all 12 unaggregated scores
Descriptive comparison and visualization	Kruskal–Wallis test; principal component analysis

3 Results and Discussion

3.1 Descriptive Statistics

Table 4 summarizes the four features. Mean scores were close to one another (80.76–81.43), and standard deviations were relatively small (2.54–3.38), indicating a compressed distribution. The lowest observed score across all assessments was 77, with 34 of the 432 individual scores (7.9%) recorded at this value. This indicates that 77 was the lowest score observed in the dataset, although it does not necessarily imply that a formal minimum score was imposed by the school's grading policy. These characteristics suggest limited score variability, which may affect the ability of clustering methods to distinguish meaningful patterns in the data. The four features were strongly and positively correlated (Pearson r = 0.67–0.86), the strongest association being between ASAS-T and ASAS-NT (r = 0.86) and the weakest between SUM and ASAS-NT (r = 0.67). Pupils who scored well on one instrument therefore tended to score well on the others, but not perfectly. Across the semester, the class mean was 1.80 points higher at TP7 (81.72) than at TP1 (79.92); because these are different assessments, this is a descriptive difference and not a measured learning gain.

Table 4. Descriptive statistics of the clustering features (N = 36).

Feature	Mean	SD	Min	Max
TP (mean of TP1–TP7)	80.76	2.54	77.29	87.86
SUM (mean of Chapters 1–3)	81.43	3.05	77.00	89.67
ASAS-T (test)	80.89	3.09	77.00	89.00
ASAS-NT (non-test)	81.08	3.38	77.00	90.00
Overall composite*	81.04	2.72	77.32	87.33

Note: *Unweighted mean of the four features; descriptive only and not used as a clustering input.

3.2 Selecting the Number of Clusters

Table 5 and Figure 2 report the validity indices for K = 2–8. The SSE fell steadily from 56.99 at K = 2 to 13.28 at K = 8 without a pronounced elbow; successive reductions were 28.1% (from 2 to 3 clusters), 25.5% (3 to 4), and 22.0% (4 to 5). The silhouette coefficient was highest at K = 2 (0.560), and the Davies–Bouldin (0.683) and Calinski–Harabasz (51.91) indices also favored K = 2. For K = 3–8, the silhouette values stayed within a narrow range (0.317–0.401) and the Davies–Bouldin values within 0.791–0.999, so the indices did not single out a clear alternative.

Table 5. Cluster validity indices for K = 2–8 (standardized four-feature data).

K	SSE	Silhouette	Davies–Bouldin	Calinski–Harabasz
2	56.99	0.560	0.683	51.91
3	40.97	0.321	0.999	41.50
4	30.51	0.317	0.989	39.68
5	23.79	0.374	0.897	39.15
6	19.15	0.367	0.900	39.11
7	15.17	0.401	0.830	41.06
8	13.28	0.384	0.791	39.38

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 4, November 2026, Page 700-709
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i4.837

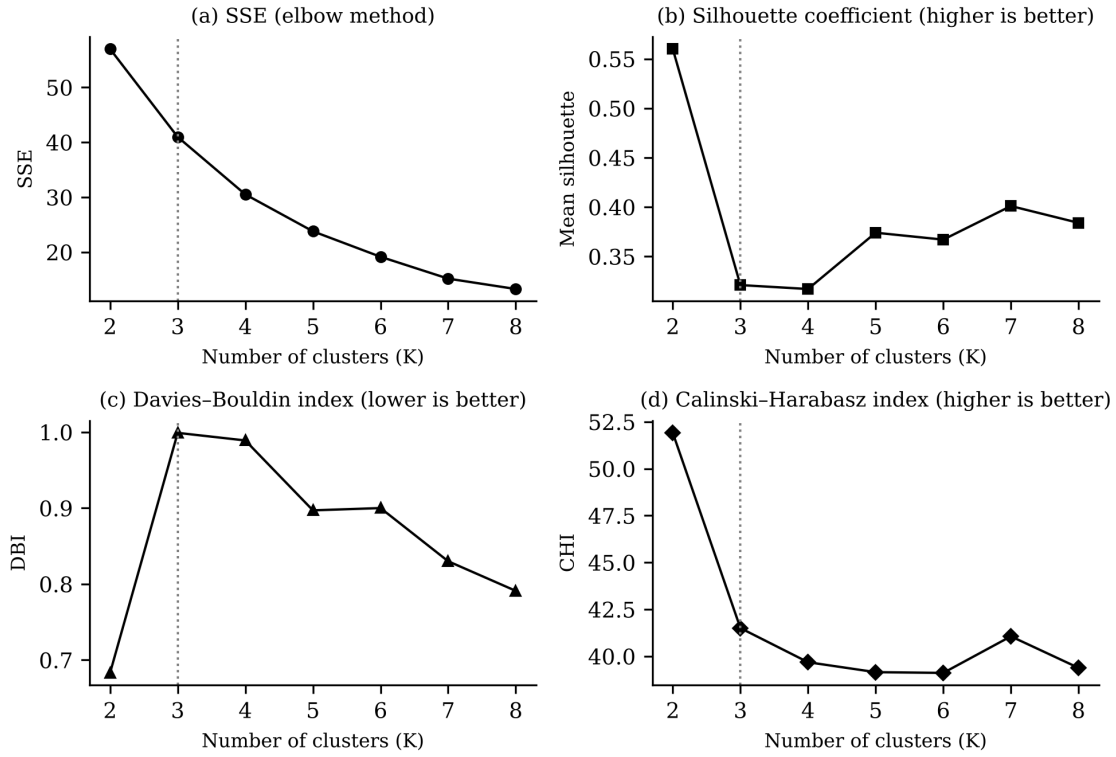

Figure 2. Cluster selection criteria for K = 2–8: (a) SSE (elbow method), (b) mean silhouette coefficient, (c) Davies–Bouldin index, and (d) Calinski–Harabasz index; the dotted line marks the retained solution (K = 3).

The two-cluster partition separates eight high-scoring pupils from 28 others. It is the statistically cleanest division but is pedagogically coarse, because it places pupils who score near the lowest observed values together with pupils who score moderately above them. Three clusters split those 28 pupils into groups of 14 and 15 and leave the high-scoring group almost unchanged (seven of its eight members). Larger values of K would produce ever smaller groups, which is hard to justify for a class of 36 pupils (at $K = 7$, the average cluster has about five members). We therefore retained $K = 3$ as an interpretive working solution, not as the statistically preferred one (Table 6). This is a judgment and not an index-driven optimum: $K = 2$ is better supported by all three internal indices, and the three-cluster solution should be read as a finer description of the same continuum. This tension illustrates the concern that validity indices alone cannot settle whether a segmentation is meaningful for teaching [15].

Table 6. Comparison of the two-cluster and three-cluster solutions.

Criterion	K = 2	K = 3
SSE	56.99	40.97
Silhouette (higher is better)	0.560	0.321
Davies–Bouldin (lower is better)	0.683	0.999
Calinski–Harabasz (higher is better)	51.91	41.50
Cluster sizes	8 and 28	7, 14, and 15
Statistical support	Stronger on all three indices	Weaker on all three indices
Reading for teaching	One high group and one undifferentiated group of 28 pupils	High, moderate, and developing groups that can be followed up differently

The difference that matters for teaching lies in the 28 pupils outside the high-scoring group. A teacher using the two-group view would plan one response for all 28, whereas the three-group view separates 14 moderate pupils (composite mean 81.11), for whom targeted feedback on specific objectives and chapters may suffice, from 15 developing pupils (composite mean 78.81), for whom a diagnostic follow-up may come first (Table 8). Whether this finer distinction is worth its weaker statistical support is a decision for the teacher; it is not an empirical finding of this study.

With 50 restarts, the three-cluster partition was identical across ten random seeds (ARI = 1.00). Single-start runs were much less stable (mean ARI = 0.54 across 100 seeds; minimum 0.34), which shows that multiple

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 4, November 2026, Page 700-709
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i4.837

restarts are necessary when groups are weakly separated. The average silhouette of the retained solution was 0.321 (0.323, 0.155, and 0.474 for C1, C2, and C3). Values in this range are commonly interpreted as weak structure [21], and the low value for C2 shows that the moderate group overlaps with its neighbors. Clustering all twelve scores instead of the four features gave a similar but not identical partition (ARI = 0.62, silhouette = 0.265), so the exact membership of boundary pupils depends on how the scores are represented.

3.3 Cluster Profiles

Table 7, Figure 3, and Figure 4 present the three clusters, ordered by composite score. The labels are neutral descriptions of one semester's scores for teacher planning; the lowest group is called "developing" rather than "basic" to avoid wording that could stigmatize pupils. Cluster C1 (high-performing; n = 7, 19.4%) had the highest means on all four features (84.57–86.86) and a composite mean of 85.68. Cluster C2 (moderate; n = 14, 38.9%) scored between 80.64 and 82.00 (composite 81.11). Cluster C3 (developing; n = 15, 41.7%) scored between 78.53 and 78.96 (composite 78.81), only slightly above the lowest observed score. Kruskal–Wallis tests indicated that the clusters differed on every feature (TP: H = 20.14; SUM: H = 23.28; ASAS-T: H = 24.69; ASAS-NT: H = 23.70; df = 2; all $p < 0.001$). As noted, these tests are descriptive only: the clusters were built from the same features, so the p-values are not external validation and are not interpreted as evidence of true group differences. The first principal component of the four standardized features explained 81.8% of the variance and loaded almost equally on all features (0.49–0.51), and the second explained 10.4%; the clusters are therefore ordered mainly along a single overall-performance axis, and a second axis separating continuous assessment (TP and SUM) from the end-of-semester assessment (ASAS) contributes little.

Table 7. Cluster profiles: mean (standard deviation) of each feature.

Cluster	n (%)	TP	SUM	ASAS-T	ASAS-NT	Comp.*
C1 High	7 (19.4)	84.57 (2.04)	85.57 (2.96)	85.71 (1.60)	86.86 (2.67)	85.68 (1.49)
C2 Moderate	14 (38.9)	80.81 (1.41)	82.00 (1.66)	81.00 (2.18)	80.64 (1.50)	81.11 (0.86)
C3 Developing	15 (41.7)	78.94 (1.32)	78.96 (1.18)	78.53 (0.83)	78.80 (0.94)	78.81 (0.90)
All pupils	36 (100.0)	80.76 (2.54)	81.43 (3.05)	80.89 (3.09)	81.08 (3.38)	81.04 (2.72)

Note: Values are mean (standard deviation). *Comp. = overall composite, the unweighted mean of the four features.

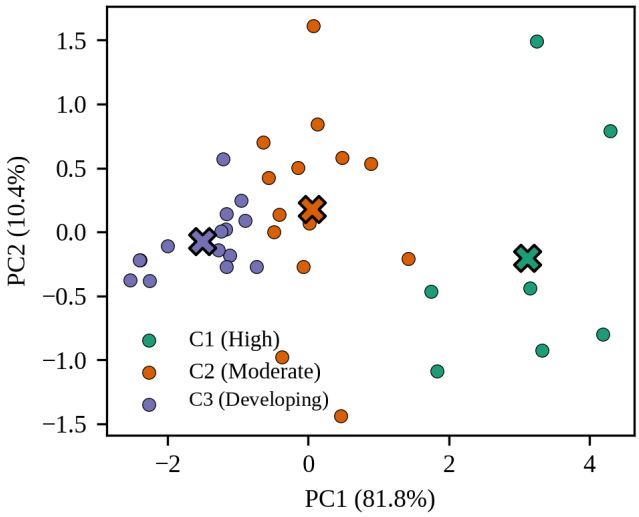

Figure 3. Pupils projected onto the first two principal components of the standardized features (PC1 + PC2 = 92.2% of variance); crosses mark cluster centroids.

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 4, November 2026, Page 700-709
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i4.837

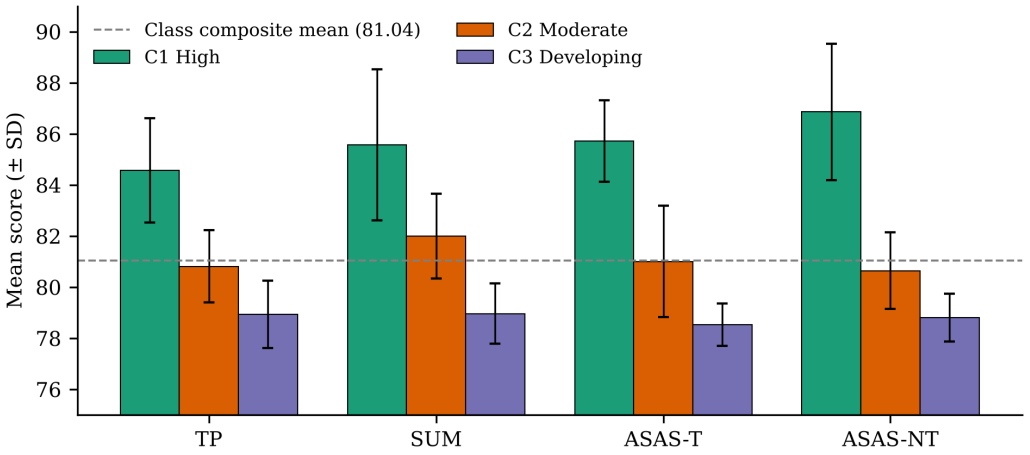

Figure 4. Mean feature scores (± SD) of the three clusters; the dashed line marks the class composite mean (81.04).

Figure 5 shows the mean score of each cluster on all 12 assessments. The order C1 > C2 > C3 held on every assessment, but the score patterns across assessments differed. Between the first and last of the seven learning-objective assessments, the C1 mean moved from 83.00 (TP1) to 86.14 (TP7), a difference of +3.14 points; the C2 mean differed by +2.29 points (79.79 to 82.07), and the C3 mean by only +0.73 points (78.60 to 79.33). The gap between C1 and C3 was therefore 4.40 points at TP1 and 6.81 points at TP7. On the chapter-level assessments, the C1 mean moved from 83.00 (Chapter 1) to 87.71 (Chapter 3), whereas the C2 (81.93 to 81.86) and C3 (79.00 to 79.00) means were essentially unchanged. These are differences between separate instruments whose difficulty and scoring were not equated, so they are descriptive score differences and not measures of within-pupil learning growth; the compressed range and the floor at 77 also limit the variation that any cluster, particularly C3, could show. The average within-pupil standard deviation across the 12 assessments was 3.04 for C1, 1.89 for C2, and 0.98 for C3, so high performers were also the most variable, and pupils in the developing cluster were almost uniformly close to 79. The difference between the non-test and test components of ASAS was small in every cluster (+1.14 for C1, −0.36 for C2, and +0.27 for C3), which is consistent with the strong correlation between the two components.

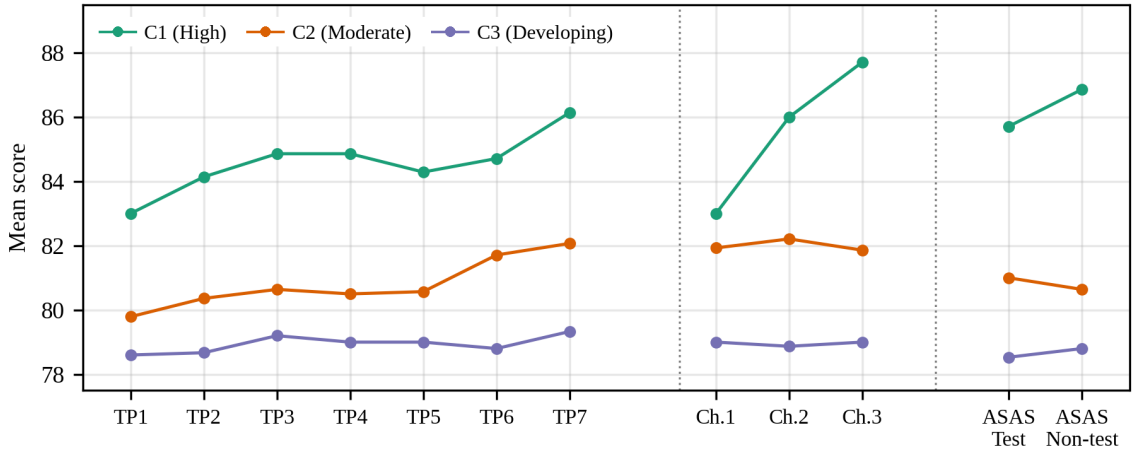

Figure 5. Mean score of each cluster on the 12 assessments (Ch. = chapter-level summative assessment; ASAS-T = test; ASAS-NT = non-test).

3.4 Discussion and Pedagogical Implications

The three-group description (high, moderate, and developing) agrees with the high, average, and low groups reported for university cohorts [10], [13], which suggests that this coarse structure is not specific to one level of education. The present data add a descriptive detail across the semester that a single final grade would have hidden: the clusters were already separated at the first learning-objective assessment, and the gap between the high-performing and developing clusters was larger at TP7 than at TP1. In this class, pupils in the higher-

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 4, November 2026, Page 700-709
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i4.837

scoring cluster showed larger differences between early and late assessments, whereas pupils who started near the lowest observed score showed almost none. Because difficulty and scoring were not equated across assessments, and because the lowest cluster sits close to the floor of 77, these differences are not evidence of learning gain.

Each profile suggests a different follow-up, which should be treated as a hypothesis for the teacher to check and not as a prescription. Table 8 links each profile to its observed evidence, a proposed teaching response, and a follow-up measure; these responses are suggestions drawn from descriptive patterns, have not been experimentally evaluated, and are not validated interventions. For C1, the higher TP7 than TP1 mean and higher variability point to a possible need for enrichment, such as extension tasks or peer-tutoring roles, so that capable pupils continue to be challenged. For C2, the unchanged chapter-level means show that these pupils scored no higher on Chapter 3 than on Chapter 1, even though their TP7 mean was modestly higher than their TP1 mean, so targeted feedback on the specific objectives and chapters where individual pupils lose points may be more useful than whole-class review [1]. For C3, the uniformly low and nearly flat scores cannot reveal why scores stayed close to the minimum. Following the logic of diagnostic assessment in the Merdeka Curriculum [2], [3], the teacher should first identify which competencies within the subject, or which circumstances, lie behind the pattern, and only then plan remedial work.

Table 8. Instructional follow-up matrix (hypotheses for the teacher to check; not validated interventions).

Profile	Observed evidence	Proposed teaching response	Follow-up measure
C1 High (n = 7)	Highest means on all four features (composite 85.68); TP7 mean 3.14 points above TP1; highest within-pupil variability (SD 3.04)	Enrichment: extension tasks or peer-tutoring roles so that capable pupils stay challenged	Check whether the variability reflects uneven mastery of particular objectives; compare with next semester's scores
C2 Moderate (n = 14)	Composite 81.11; TP7 mean 2.29 points above TP1, but chapter-level means unchanged (81.93 and 81.86); weakest silhouette (0.155), so boundary pupils may move between clusters	Targeted feedback on the objectives and chapters where individual pupils lose points, rather than whole-class review	Item-level review of chapter assessments; a short re-check on the identified objectives
C3 Developing (n = 15)	Composite 78.81, close to the observed minimum of 77; TP7 mean 0.73 points above TP1; chapter-level means unchanged (79.00); within-pupil SD 0.98	Diagnostic follow-up first, to identify the competencies or circumstances behind the pattern, and only then remedial work	Competency-specific diagnostic task and teacher observation; review of the pattern at the next assessment

These findings also illustrate the interpretive gap identified in the literature [15]. The algorithm can group pupils, but it cannot explain the groups; that explanation depends on the teacher's knowledge of the pupils and on the learning goals. Computing and artificial intelligence technologies are increasingly discussed as tools for education [27], including early-warning systems that predict student academic risk [28], but in a setting like this one they remain aids to, not substitutes for, the teacher's judgment. Because the pupils are minors and the scores are context-dependent [14], cluster labels should stay with teachers as planning tools, should not be communicated to pupils or parents as fixed categories, and should be recomputed every semester because pupils can move between clusters [11].

3.5 Limitations

Several limitations should be kept in mind. First, the data come from one class of 36 pupils in one semester, so the results describe this class and are not generalizable. Second, scores are compressed into a narrow band with a floor at 77, which reduces the separation between pupils and probably explains the weak silhouette values; the clusters are better seen as segments of a continuum than as natural groups. It is also not established whether 77 is a formal reporting floor, and no rank-based or alternative normalization sensitivity check was performed; the only sensitivity analysis used all twelve unaggregated scores. Third, only assessment scores were available. The dataset has no scores for separate competencies within the subject, no attendance or background information, and no information on how each assessment was scored, so the causes of the patterns cannot be inferred. The instruments are also not shown to be comparable in difficulty or scoring, so differences between early and late assessments cannot be read as learning gains. Fourth, the choice of three clusters involved judgment, and the partition changed moderately when all twelve scores were used. Fifth, K-means favors compact, similarly sized groups and uses Euclidean distance [8], so other methods, such as hierarchical clustering or Gaussian mixture models, could produce different groupings. Finally, the instructional suggestions in Section 3.4 are hypotheses

derived from descriptive patterns; they have not been experimentally evaluated and are not validated interventions.

4 Conclusion

This study showed how K-means clustering of routinely collected multi-assessment scores can describe learning performance patterns in a primary school Mekarsi class. Four standardized features built from learning-objective, chapter-level, and end-of-semester assessments yielded three interpretable groups: high-performing (19.4%), moderate (38.9%), and developing pupils (41.7%). The groups differed not only in level but also in the difference between the first and last learning-objective scores, which was largest for the high-performing group and smallest for the developing group; because the assessments are not shown to be comparable, this difference is descriptive and not evidence of learning gain. The validity indices supported a two-cluster solution more strongly than the three-cluster solution, and the average silhouette of 0.321 indicates weak structure, so the clusters should be used as a cautious planning aid and not as fixed categories.

Future work should replicate the analysis with several classes and semesters, include skill-specific scores and contextual variables, compare K-means with alternative clustering methods and normalizations, and combine the quantitative profiles with teacher or pupil interviews to test whether the interpretations proposed here match classroom reality. Any instructional response suggested here should be evaluated before it is treated as effective.

Disclosure Statements

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the research, authorship, and publication of this article.

Author Contributions

R.K.: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing. S.S.: Investigation, Data Collection, Validation, Writing – Review & Editing. All authors have read and approved the final version of the manuscript.

Acknowledgements

The authors would like to acknowledge SD Negeri No. 104222 Sei Glugur, Deli Serdang, for providing access to the anonymized assessment records used in this study.

Data Availability Statement

The anonymized score data analyzed in this study are available from the corresponding author upon reasonable request, subject to permission from the school.

Ethics Statement

This study analyzed existing, anonymized assessment records provided by the school; pupil names were replaced with codes before analysis, no additional data were collected from pupils, and cluster labels were used only as descriptive planning summaries for the teacher.

References

[1] P. Black and D. Wiliam, “Assessment and classroom learning,” *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7–74, 1998, doi: 10.1080/0969595980050102.

[2] D. Amelia, I. Mutmainah, I. Indriyani, and N. S. Carmenette, “The study of assessment in the Merdeka curriculum at elementary school,” *Indonesian Journal of Primary Education*, vol. 7, no. 2, pp. 121–128, 2023, doi: 10.17509/ijpe.v7i2.57343.

[3] A. R. Tsalsabilla and W. Oktavia, “English teachers’ perception on assessment changes in the Merdeka curriculum at SMAN 7 Padang,” *Journal of English Language Teaching*, vol. 13, no. 2, pp. 622–636, 2024. [Online]. Available: https://ejournal.unp.ac.id/index.php/jelt/article/view/128638

[4] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[5] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, Art. no. e1355, 2020, doi: 10.1002/widm.1355.

[6] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017, doi: 10.1109/ACCESS.2017.2654247.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281–297.

[8] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.

[9] S. J. Alalawi, I. N. Mohd Shaharanee, and J. Mohd Jamil, "Clustering student performance data using k-means algorithms," *Journal of Computational Innovation and Analytics*, vol. 2, no. 1, pp. 41–55, 2023, doi: 10.32890/jcia2023.2.1.3.

[10] N. I. Mohd Talib, N. A. Abd Majid, and S. Sahran, "Identification of student behavioral patterns in higher education using K-means clustering and support vector machine," *Applied Sciences*, vol. 13, no. 5, Art. no. 3267, 2023, doi: 10.3390/app13053267.

[11] E. Tuyishimire, W. Mabuto, P. Gatabazi, and S. Bayisingize, "Detecting learning patterns in tertiary education using K-means clustering," *Information*, vol. 13, no. 2, Art. no. 94, 2022, doi: 10.3390/info13020094.

[12] M. Ahmad, N. I. Arshad, and A. Sarlan, "An analysis of students' academic performance using K-means clustering algorithm," in *Advances on Intelligent Informatics and Computing, Lecture Notes on Data Engineering and Communications Technologies, vol. 127*. Cham, Switzerland: Springer, 2022, pp. 309–318, doi: 10.1007/978-3-030-98741-1_26.

[13] A. S. Abdelmagid and A. I. M. Qahmash, "Utilizing the educational data mining techniques 'Orange Technology' for detecting patterns and predicting academic performance of university students," *Information Sciences Letters*, vol. 12, no. 3, pp. 1415–1431, 2023, doi: 10.18576/isl/120330.

[14] Irmayanti, H. F. Nasution, and Rohani, "Clustering elementary school students' academic achievement profiles using K-means, the elbow method, and silhouette evaluation," *Information Technology Education Journal*, vol. 5, no. 3, pp. 261–274, 2026, doi: 10.59562/intec.v5i3.13453.

[15] F. R. Hilyawan, M. Syafrullah, and I. Sudija, "A systematic literature review on the optimization of K-means and agglomerative clustering for student performance segmentation: A comparative analysis of elbow and silhouette methods," *Edcomtech: Jurnal Kajian Teknologi Pendidikan*, vol. 10, no. 1, pp. 73–88, 2025, doi: 10.17977/um039v10i12025p73-88.

[16] A. Punhani, N. Faujdar, K. K. Mishra, and M. Subramanian, "Binning-based silhouette approach to find the optimal cluster using K-means," *IEEE Access*, vol. 10, pp. 115025–115032, 2022, doi: 10.1109/ACCESS.2022.3215568.

[17] A. Rykov, R. C. de Amorim, V. Makarenkov, and B. Mirkin, "Inertia-based indices to determine the number of clusters in K-means: An experimental evaluation," *IEEE Access*, vol. 12, pp. 11761–11773, 2024, doi: 10.1109/ACCESS.2024.3350791.

[18] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.

[19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[21] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[22] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[23] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[24] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985, doi: 10.1007/BF01908075.

[25] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952, doi: 10.1080/01621459.1952.10483441.

[26] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, Art. no. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

[27] R. Kaban et al., *Teknologi Komputasi dan Kecerdasan Buatan dalam Pendidikan [Computing Technology and Artificial Intelligence in Education]*, 1st ed. Medan, Indonesia: Yayasan Kita Menulis, 2026.

[28] R. Kaban, J. B. Sallim, S. Susmanto, and R. B. Mohamed, "Early warning system for student academic risk prediction using gradient boosting with SMOTE-based class balancing," *JCEIT: Journal of Computer Engineering and Information Technology*, vol. 2, no. 3, 2026.