Unsupervised Clustering and Visual Learning Analytics for Intelligent Competency Assessment in Government Education Selection Systems

Chen Junru[1,*], Cheng Fan[2]

[1]Liaoning University of Science and Technology, China

[2]Southwest University, China

ABSTRACT

Artificial intelligence and learning analytics have become increasingly important in supporting data driven educational assessment and competency profiling. This study proposes an unsupervised machine learning approach for competency profiling of government education selection participants using clustering and learning analytics techniques based on Seleksi Kompetensi Dasar (SKD) assessment data. The dataset consisted of 2,150 participants and included four assessment variables, namely Tes Wawasan Kebangsaan (TWK), Tes Intelegensia Umum (TIU), Tes Karakteristik Pribadi (TKP), and total SKD score. Prior to clustering, all variables were standardized using the StandardScaler method. The K Means algorithm was employed to identify latent competency structures, while the optimal number of clusters was determined using the Elbow Method and Silhouette Score analysis. Clustering performance was evaluated using Silhouette Score, Davies Bouldin Index, and Calinski Harabasz Score. The experimental results demonstrated that the optimal clustering configuration was achieved at k=2with a silhouette score of 0.585, indicating good cluster cohesion and separation. The findings revealed two distinct participant competency groups, where Cluster 0 represented high performing participants with consistently higher scores across all SKD dimensions, while Cluster 1 demonstrated lower competency readiness. Among all variables, TIU emerged as the most discriminative factor in competency differentiation. Furthermore, PCA and t SNE visualizations confirmed clear cluster separation with minimal overlap, indicating the presence of meaningful latent competency structures within the dataset. These findings demonstrate that unsupervised machine learning and learning analytics approaches can effectively support intelligent educational assessment, competency mapping, and data driven decision making in educational selection systems.

Keywords Unsupervised Machine Learning, Learning Analytics, Competency Profiling, Educational Data Mining, Intelligent Educational Assessment

Submitted 15 June 2026
Accepted 25 August 2026
Published 17 November 2026

*Corresponding author
Cheng Junru,
315780275@qq.com

Additional Information and Declarations can be found on page 304

DOI: 10.63913/ail.v2i4.75
© Copyright
2026 Junru and Fan

Distributed under
Creative Commons CC-BY 4.0

Introduction

The rapid advancement of artificial intelligence and data driven technologies has significantly transformed the educational sector, particularly in the areas of learning analytics, educational data mining, and intelligent assessment systems [1], [2]. Educational institutions increasingly utilize machine learning techniques to analyze large scale educational data in order to identify learning patterns, evaluate participant performance, and support evidence-based decision making. In recent years, unsupervised machine learning approaches have attracted considerable attention because of their ability to discover hidden structures and latent patterns within educational datasets without requiring predefined labels. Among these approaches, clustering techniques have been

How to cite this article: C. Junru and C. Fan, "Unsupervised Clustering and Visual Learning Analytics for Intelligent Competency Assessment in Government Education Selection Systems," *Artif. Intell. Learn.*, vol. 2, no. 4, pp. 288-307, 2026.

widely applied for student grouping, learning behavior analysis, academic performance prediction, and competency profiling [3]. The integration of clustering and learning analytics provides opportunities to develop more adaptive, explainable, and data driven educational assessment frameworks.

Competency profiling has become an important issue in educational selection and assessment systems because competency characteristics often vary significantly among participants. In government education selection environments, identifying participant competency patterns is essential for supporting fair assessment, educational placement, and competency mapping processes. The Seleksi Kompetensi Dasar (SKD) assessment is designed to evaluate participant competencies through three primary dimensions, namely Tes Wawasan Kebangsaan (TWK), Tes Intelegensia Umum (TIU), and Tes Karakteristik Pribadi (TKP). These dimensions collectively represent civic knowledge, analytical reasoning, and behavioral competency. Although SKD assessment results contain valuable educational information, most existing evaluation processes still rely on conventional descriptive analysis and threshold-based assessment approaches. Such approaches are limited in their ability to identify hidden competency structures, latent participant characteristics, and multidimensional competency relationships within assessment data.

Several previous studies have applied clustering techniques in educational data mining for student performance analysis, learning behavior classification, and academic competency grouping [4]. K Means clustering has frequently been utilized because of its computational efficiency and ability to generate interpretable group structures. Other studies have employed learning analytics visualization techniques such as Principal Component Analysis (PCA), t Distributed Stochastic Neighbor Embedding (t SNE), heatmaps, and dashboard-based analytics to improve the interpretability of educational data patterns. Recent developments in artificial intelligence in education have also emphasized the importance of explainable and interpretable machine learning models to support intelligent educational assessment and decision support systems [5]. However, most existing studies primarily focus on academic learning management systems, online learning environments, or university student performance datasets. Research involving competency profiling in government education selection systems remains relatively limited.

In addition, previous educational clustering studies commonly focus only on basic clustering implementation without comprehensive competency interpretation, visual learning analytics integration, or multi perspective cluster validation. Many studies also lack explainable competency profiling frameworks capable of translating clustering results into interpretable educational insights for assessment and decision-making purposes. Furthermore, limited research has explored the integration of clustering, visual analytics, and competency mapping within SKD based government education selection contexts. As a result, there remains a research gap regarding how unsupervised machine learning can be effectively utilized to identify latent competency structures and support intelligent educational assessment in government education selection systems.

To address these limitations, this study proposes an unsupervised machine learning framework for competency profiling using clustering and learning analytics approaches based on SKD assessment data. The proposed

framework utilizes K Means clustering to identify participant competency groups using TWK, TIU, TKP, and total SKD score variables. The clustering process is evaluated using Silhouette Score, Davies Bouldin Index, and Calinski Harabasz Score to ensure clustering quality and stability. In addition, visual learning analytics techniques including PCA, t SNE, heatmap visualization, boxplot analysis, and hierarchical clustering are employed to improve interpretability and support competency analysis. Through this approach, the study aims to identify latent participant competency structures, analyze competency characteristics across clusters, and provide interpretable educational insights that support intelligent educational assessment and competency mapping.

The main contributions of this study are threefold. First, this study proposes an artificial intelligence-based competency profiling framework for government education selection using unsupervised machine learning techniques. Second, the study integrates clustering with visual learning analytics to improve the interpretability and explainability of participant competency structures. Third, the study identifies the dominant role of cognitive and analytical competency, particularly the TIU component, in differentiating participant competency profiles. The findings of this study are expected to contribute to the development of intelligent educational assessment systems, learning analytics applications, and data driven decision making frameworks in educational selection environments.

Literature Review and Related Works

Artificial intelligence and machine learning have become increasingly important in educational environments due to their ability to analyze complex educational data and support intelligent decision-making processes. In recent years, educational data mining and learning analytics have been widely utilized to identify hidden learning patterns, evaluate student performance, and improve educational assessment systems [6], [7]. The integration of machine learning techniques into educational systems enables the development of adaptive, data driven, and interpretable educational frameworks capable of supporting competency evaluation and learning behavior analysis [8].

Unsupervised machine learning techniques, particularly clustering algorithms, have been extensively applied in educational research to identify latent group structures and classify learners based on academic performance, behavioral patterns, and competency characteristics [9], [10]. Among various clustering methods, K Means remains one of the most frequently used algorithms because of its computational efficiency, scalability, and interpretability in educational datasets [11]. Previous studies have demonstrated that clustering techniques can effectively group students according to learning styles, academic achievement, and engagement behavior, thereby supporting personalized educational interventions and adaptive learning systems [12], [13].

Learning analytics visualization has also become an essential component in educational data interpretation. Techniques such as Principal Component Analysis (PCA), t Distributed Stochastic Neighbor Embedding (t SNE), heatmaps, and dashboard based visual analytics have been widely adopted to improve the interpretability of educational patterns and learner behaviors [14], [15]. Visual learning analytics enables educators and decision makers to understand complex multidimensional relationships within educational datasets and facilitates more explainable educational assessment processes [16].

Competency profiling represents another important area within educational analytics and intelligent assessment systems. Several studies have explored the use of machine learning for competency mapping, talent identification, and performance evaluation in academic and professional education environments [17], [18]. These approaches aim to identify participant strengths, weaknesses, and readiness levels based on multidimensional assessment indicators. In government education and recruitment systems, competency profiling plays a critical role in ensuring objective and data driven participant evaluation processes [19].

Despite the growing adoption of clustering and learning analytics in education, several limitations remain in existing studies. Most previous research primarily focuses on online learning systems, university student performance, or learning management system data. Research involving competency profiling using government educational selection datasets remains relatively limited. In addition, many existing clustering studies emphasize clustering accuracy without integrating explainable learning analytics visualization and educational interpretation frameworks [20]. Consequently, there is still limited research investigating how unsupervised machine learning and visual learning analytics can be integrated to support intelligent competency profiling and educational assessment in government education selection systems. Therefore, this study proposes a clustering-based competency profiling framework using SKD assessment data to support explainable learning analytics and intelligent educational assessment.

Methodology

Research Framework

This study employed an unsupervised machine learning framework to perform competency profiling of government education selection participants using clustering and learning analytics approaches based on SKD assessment data. The proposed framework was designed to identify latent competency structures among participants by analyzing similarities in their assessment performance across multiple competency dimensions. The overall research process consisted of several sequential stages, including dataset collection, data preprocessing, feature standardization, clustering analysis, cluster evaluation, competency profiling, visual learning analytics, hierarchical validation, and outlier detection. Figure 1 illustrates the overall workflow of the proposed research framework, beginning with the acquisition of SKD assessment data and continuing through preprocessing and clustering stages until competency interpretation and learning analytics visualization were obtained. The integration of clustering techniques and visual analytics was intended to improve both the interpretability and explainability of participant competency patterns within the educational assessment process.

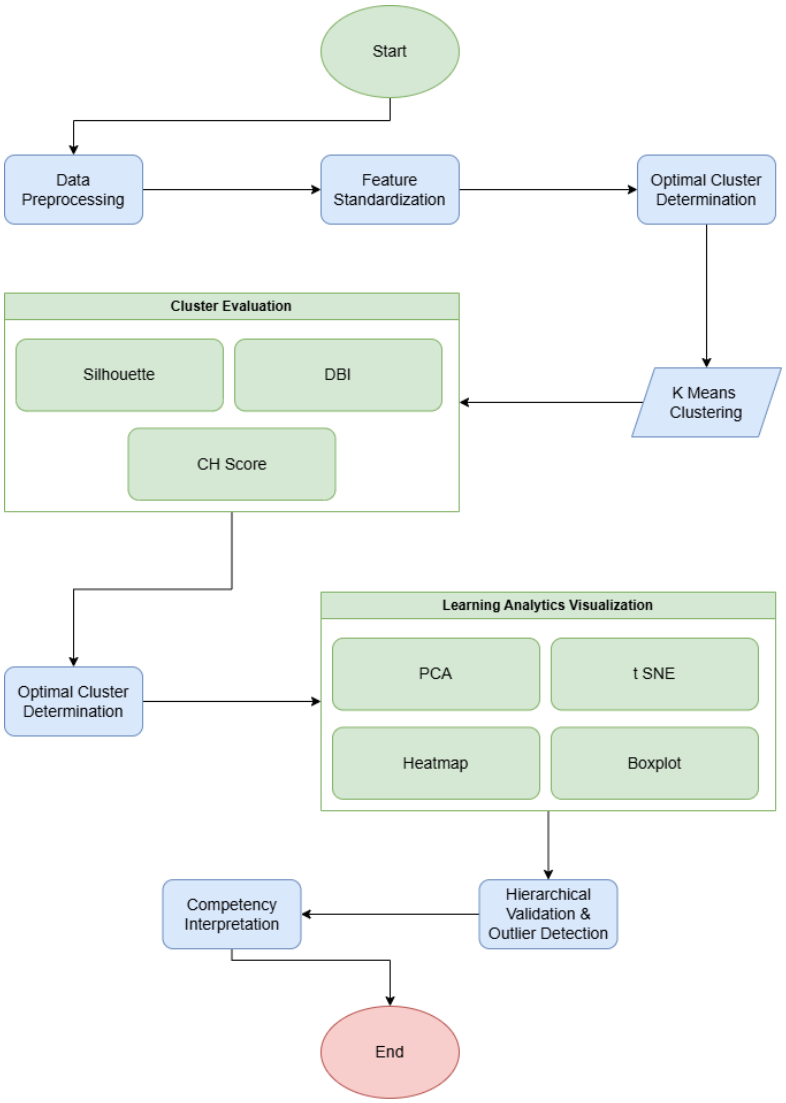


Figure 1 Research Framework

The primary objective of the proposed framework was to support intelligent educational assessment by automatically identifying participant competency groups based on SKD performance patterns. Clustering analysis was used to group participants with similar competency characteristics, while learning analytics visualization techniques were utilized to reveal hidden relationships, multidimensional competency structures, and score distribution patterns. In addition, hierarchical clustering and DBSCAN based outlier detection were incorporated to validate the clustering structure and identify anomalous participant profiles. Through this integrated framework, the study aimed to develop a more explainable and data driven competency profiling approach capable of supporting educational decision making and competency mapping within government education selection systems.

Dataset Description

The dataset used in this study consisted of SKD assessment results collected from government education selection participants in 2024. The dataset contained 2,150 participant records and 11 attributes representing participant

information and assessment results. The competency related variables used for clustering analysis included Tes Wawasan Kebangsaan (TWK), Tes Intelegensia Umum (TIU), Tes Karakteristik Pribadi (TKP), and total SKD score.

TWK represents participant competency in civic and national knowledge, TIU measures analytical and cognitive ability, while TKP evaluates behavioral and personality related competency aspects. The total SKD score represents the cumulative competency performance obtained from all assessment dimensions. These variables were selected because they collectively describe participant competency readiness from cognitive, behavioral, and civic perspectives.

The features utilized for clustering analysis are presented in table 1.

Table 1 Features Used for Clustering

Variable	Description
TWK	Tes Wawasan Kebangsaan score
TIU	Tes Intelegensia Umum score
TKP	Tes Karakteristik Pribadi score
Total Score	Total SKD assessment score

Data Preprocessing

Prior to clustering analysis, several preprocessing steps were conducted to ensure data quality, consistency, and compatibility with machine learning algorithms. Initially, all selected clustering variables were converted into numerical format using the pandas numerical conversion method to prevent invalid data types during clustering computation. Records containing missing values within the primary competency variables were removed to avoid incomplete observations that could negatively affect clustering performance and distance calculations.

After the cleaning process, feature standardization was performed using the StandardScaler technique. Standardization was necessary because clustering algorithms based on Euclidean distance are sensitive to differences in feature magnitude and variance. Without scaling, variables with larger numerical ranges could dominate the clustering process and bias centroid formation.

The standardization process transforms each feature into a standardized distribution with zero mean and unit variance using the following equation:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

$x$denotes the original feature value, μrepresents the mean of the feature, and σindicates the standard deviation. The standardized dataset was then utilized as the input for all clustering and visualization processes.

Clustering Method

This study employed the K Means clustering algorithm to identify latent competency groups among participants. K Means is an unsupervised machine learning technique that partitions observations into $k$clusters based on similarity and minimizes the distance between data points and their corresponding cluster centroids. The algorithm iteratively updates cluster centroids until convergence is achieved.

The objective function of K Means clustering is expressed as:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} || x_j - c_i ||^2 \quad (2)$$

$C_i$represents the i-th cluster and $c_i$denotes the centroid of cluster i.

To determine the optimal number of clusters, the Elbow Method and Silhouette Score analysis were performed by evaluating cluster configurations ranging from:

$$k = 2 \text{ to } 10 \quad (3)$$

The Elbow Method was used to analyze inertia reduction patterns, while the Silhouette Score was utilized to evaluate cluster cohesion and separation quality. The clustering implementation was conducted using the Scikit Learn library with a fixed random state of 42 to ensure reproducibility and consistency of experimental results.

Clustering Evaluation

To evaluate clustering quality and validate the effectiveness of the generated competency groups, three clustering evaluation metrics were employed, namely Silhouette Score, Davies Bouldin Index, and Calinski Harabasz Score.

The Silhouette Score measures the similarity of data points within the same cluster relative to data points in neighboring clusters. Higher silhouette values indicate better clustering performance with stronger separation and cohesion.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$

$a(i)$denotes the average intra cluster distance and $b(i)$represents the minimum average distance to other clusters.

The Davies Bouldin Index evaluates cluster similarity and separation based on centroid distance and cluster dispersion. Lower DBI values indicate better clustering quality.

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (5)$$

$S_i$and $S_j$represent cluster dispersion and M_{ij}denotes the centroid distance between clusters.

The Calinski Harabasz Score measures the ratio between between cluster dispersion and within cluster dispersion.

$$CH = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)} \quad (6)$$

Higher Calinski Harabasz values indicate more compact and better separated clusters.

Competency Profiling and Learning Analytics

After clustering, competency profiling was conducted by calculating the mean score of each SKD variable within every cluster. This process enabled the

identification of dominant competency characteristics and participant readiness patterns. Each cluster was interpreted based on TWK, TIU, TKP, and total score distributions to determine differences in cognitive, civic, and behavioral competency structures.

To improve interpretability and support learning analytics analysis, several visualization techniques were applied, including Principal Component Analysis (PCA), t Distributed Stochastic Neighbor Embedding (t SNE), heatmap visualization, boxplot analysis, and radar chart profiling. PCA and t SNE were utilized to visualize participant distributions and cluster separation in lower dimensional space, while heatmaps and radar charts were employed to compare competency intensity across clusters. In addition, boxplots were used to analyze score variability, distribution spread, median values, and outlier patterns within each competency group.

Hierarchical Validation and Outlier Detection

To validate the clustering structure, hierarchical clustering using Ward linkage was performed and visualized through a dendrogram representation. Hierarchical clustering enables the identification of nested participant relationships and hierarchical competency structures within the dataset. The consistency between hierarchical clustering and K Means clustering results was used to confirm clustering stability and robustness.

Furthermore, DBSCAN was employed to identify anomalous participant profiles and detect outlier competency patterns within the dataset. DBSCAN classifies isolated observations as noise when they do not belong to sufficiently dense regions. Participants assigned with a cluster label of:

$$-1 \tag{7}$$

were categorized as anomalous competency profiles. The integration of clustering, learning analytics visualization, hierarchical validation, and outlier detection enabled the development of a comprehensive artificial intelligence based competency profiling framework for intelligent educational assessment and educational decision support systems.

Results

Clustering Performance Evaluation

The clustering process was performed using the K-Means algorithm on four SKD assessment variables, namely TWK, TIU, TKP, and total score, after all features were standardized using the StandardScaler technique to eliminate scale differences among variables. To identify the optimal number of clusters, the Elbow Method and Silhouette Score analysis were applied by testing cluster configurations ranging from $k = 2$to $k = 10$. Figure 2 presents the Elbow Method results, where the inertia value decreased sharply between $k = 2$and $k = 4$, indicating that the data structure begins to stabilize within this interval. The most significant reduction occurred at $k = 2$, suggesting that the dataset naturally forms two dominant competency groups. This finding was further supported by the silhouette analysis, which produced the highest silhouette coefficient at $k = 2$. The combination of these results indicates that the selected clustering configuration provides the best balance between cluster compactness and inter-cluster separation, demonstrating the suitability of K-Means clustering for

competency profiling in SKD assessment data.

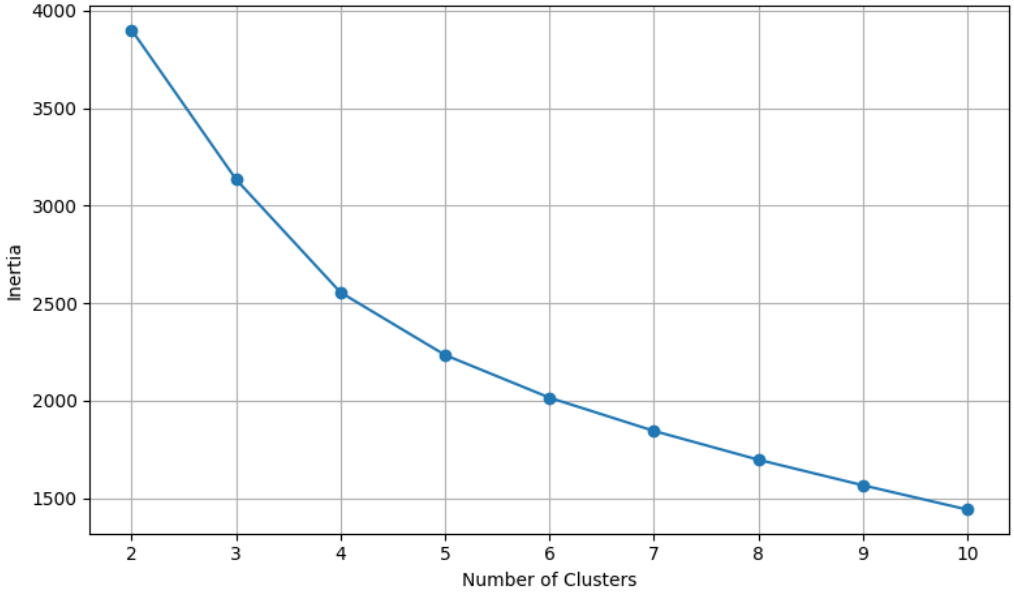


Figure 2 Elbow Method for Optimal Cluster Determination

The silhouette analysis demonstrated that the highest silhouette coefficient was achieved at:

$$k = 2 \tag{8}$$

as illustrated in figure 3. The silhouette score measures the degree of similarity between data points within the same cluster compared to those in other clusters, where higher values indicate better cluster cohesion and separation. Among all tested cluster configurations ranging from $k = 2$to $k = 10$, the configuration with two clusters produced the highest silhouette value of 0.585, indicating a strong clustering structure within the dataset. This result suggests that participants within the same cluster share similar competency characteristics, while the separation between clusters remains sufficiently distinct. The relatively high silhouette coefficient confirms that the selected clustering configuration effectively captures the underlying competency patterns of SKD participants and provides a reliable foundation for competency profiling and learning analytics interpretation.

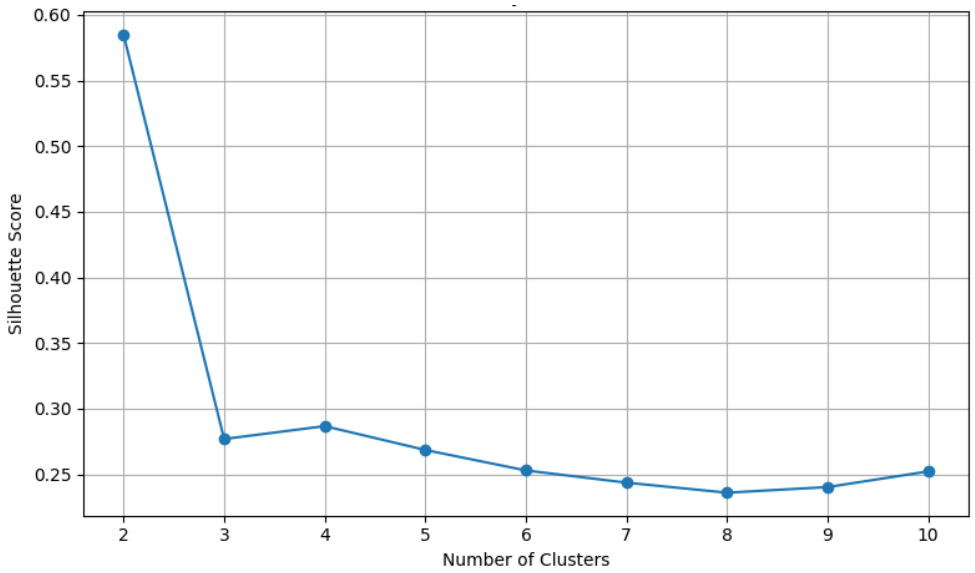


Figure 3 Silhouette Score by Number of Clusters

The clustering evaluation metrics are summarized in table 2. The clustering model achieved a silhouette score of 0.585, indicating that the generated clusters possess good internal cohesion and sufficient separation between groups. In clustering analysis, silhouette values above 0.5 generally represent a strong and reliable clustering structure, suggesting that participants within the same cluster share similar competency characteristics while remaining distinct from participants in other clusters. In addition, the Davies-Bouldin Index value of 0.683 indicates low similarity between clusters, since smaller Davies-Bouldin values reflect better inter-cluster separation and reduced overlap among groups. The clustering performance was further strengthened by the high Calinski-Harabasz Score of 2588.239, which demonstrates that the clusters are both compact and well-distributed within the feature space. Collectively, these evaluation results confirm that the K-Means algorithm successfully identified meaningful competency structures within the SKD assessment dataset and produced stable clustering outcomes suitable for competency profiling and learning analytics applications.

Table 2 Clustering Evaluation Metrics

Metric	Value
Silhouette Score	0.585
Davies-Bouldin Index	0.683
Calinski-Harabasz Score	2588.239

Cluster Distribution and Competency Profiling

The clustering model divided a total of 2,150 participants into two distinct competency groups based on similarities in their SKD assessment scores. Cluster 0 contained 1,796 participants, representing approximately 83.5% of the dataset, while Cluster 1 consisted of 354 participants or approximately 16.5% of all participants. This distribution indicates that the majority of participants belong to a higher competency category, whereas a smaller proportion demonstrates relatively lower competency characteristics. To understand the competency patterns within each group, cluster profiling was performed by calculating the

mean score of each SKD component, including TWK, TIU, TKP, and total score, for every cluster. The profiling process enables the identification of dominant competency characteristics and provides a clearer interpretation of the latent competency structure generated by the clustering model. The detailed competency profile of each cluster is presented in table 3.

Table 3 Competency Profile of Each Cluster

Cluster	TWK	TIU	TKP	Total Score
Cluster 0	91.12	138.83	195.75	425.71
Cluster 1	65.93	71.98	175.05	312.96

The results reveal substantial differences between clusters across all competency dimensions, indicating the presence of distinct participant competency profiles within the SKD assessment dataset. Cluster 0 consistently achieved higher average scores in TWK, TIU, TKP, and total score compared to Cluster 1, demonstrating stronger overall competency performance across cognitive, civic, and behavioral dimensions. In particular, the TIU component exhibited the largest score gap between clusters, with Cluster 0 achieving a significantly higher average score than Cluster 1. This finding suggests that analytical reasoning and cognitive ability serve as the primary differentiating factors in the competency profiling process. The relatively large disparity in TIU scores also indicates that intellectual capability plays a critical role in determining overall participant readiness in government education selection. Furthermore, the consistent differences across all variables confirm that the clustering model successfully captured meaningful latent competency structures within the dataset. The participant distribution across clusters is illustrated in figure 4.

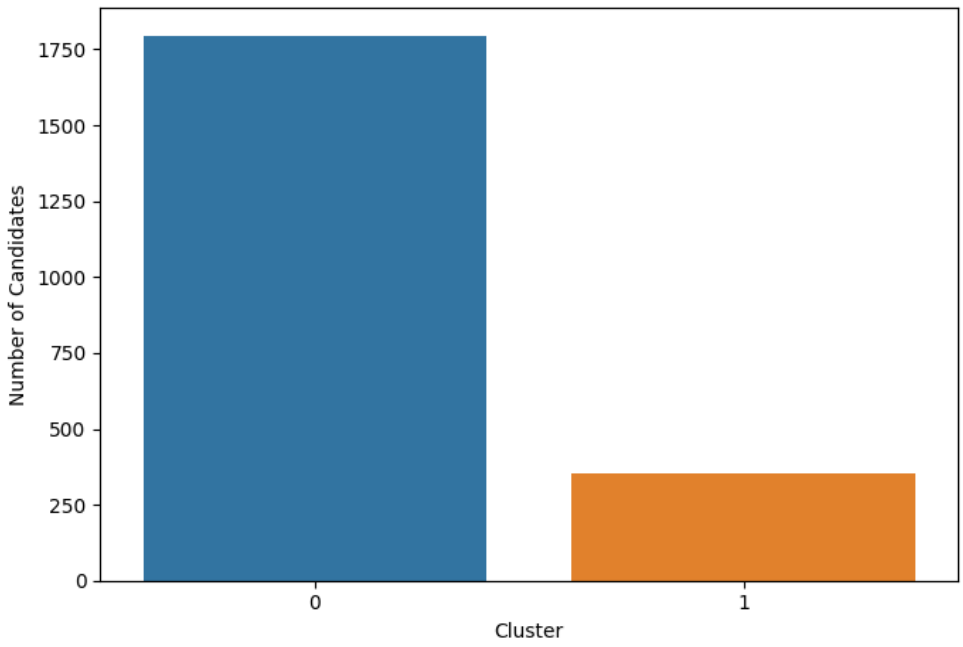


Figure 4 Distribution of Candidate Clusters

To further examine the competency characteristics of each cluster, a heatmap visualization was generated as shown in figure 5. The heatmap provides a comparative representation of the average competency intensity across all SKD

dimensions, including TWK, TIU, TKP, and total score. The visualization clearly demonstrates that Cluster 0 consistently exhibits higher competency values across all assessment variables, indicating a stronger overall competency profile. In contrast, Cluster 1 shows substantially lower performance levels in every competency dimension. Among all variables, the contrast in TIU scores appears to be the most prominent, further reinforcing the role of cognitive and analytical ability as the primary differentiating factor between clusters. The consistent pattern observed in the heatmap confirms the stability of the clustering structure and supports the effectiveness of the proposed unsupervised learning approach in identifying latent participant competency profiles.

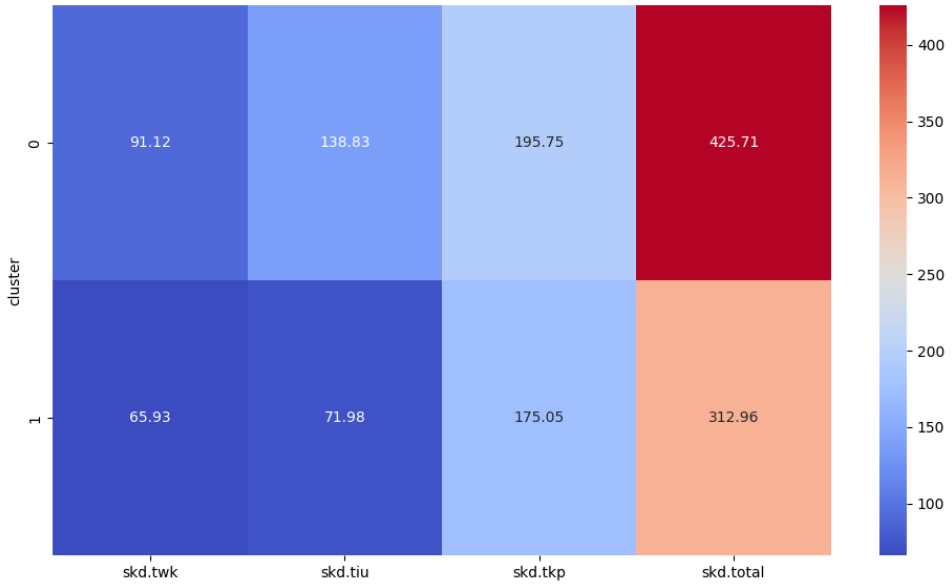


Figure 5 Cluster Profile Heatmap

Cluster Visualization

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the SKD assessment dataset and visualize the clustering structure in two-dimensional space. By transforming the original variables into principal components, PCA enables the identification of dominant variance patterns while preserving the overall structure of the data. As illustrated in figure 6, the PCA visualization demonstrates a clear separation between the two identified clusters with only minimal overlap between participant groups. Cluster 0 is primarily concentrated in the positive region of the first principal component, whereas Cluster 1 is predominantly distributed in the negative region. This separation indicates that the SKD assessment variables contain strong discriminative information that effectively differentiates participant competency profiles. The visualization also confirms that the clustering model successfully captured meaningful latent structures within the dataset, supporting the validity and interpretability of the competency profiling process.

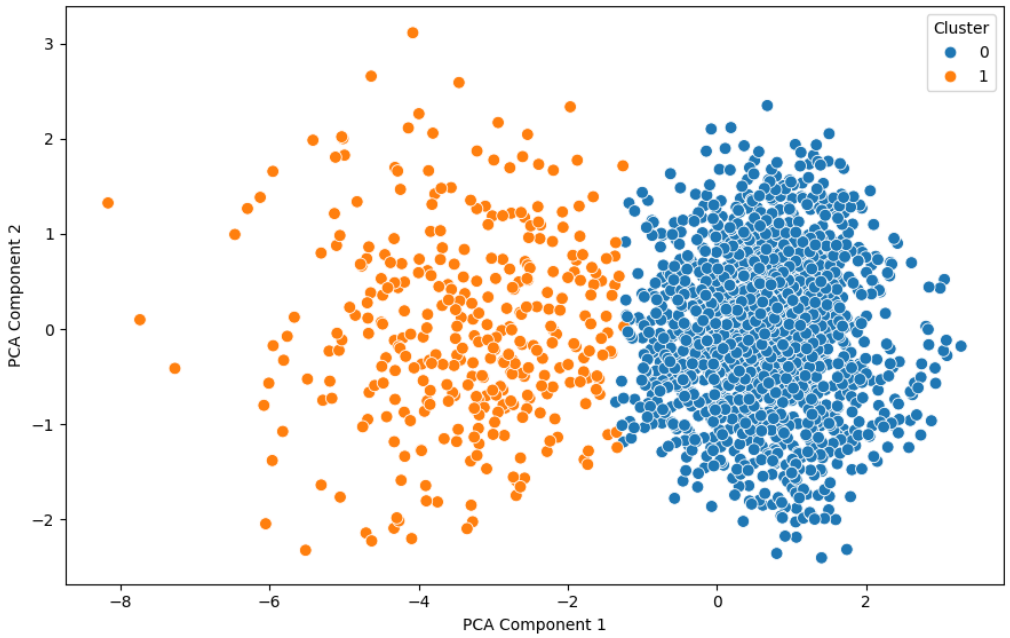


Figure 6 PCA Visualization of Candidate Clusters

To further explore nonlinear relationships and hidden data structures within the dataset, t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization was performed. Unlike PCA, which primarily captures linear variance, t-SNE is capable of preserving local neighborhood relationships and revealing complex nonlinear patterns among participants. As illustrated in figure 7, the t-SNE representation generated more compact, dense, and distinguishable cluster formations with clearer boundaries between groups. The visualization demonstrates that participants belonging to the same cluster are positioned closely together, while the separation between clusters remains highly distinct. This result indicates the presence of latent competency structures within the SKD assessment dataset and further confirms that the proposed clustering approach successfully identified meaningful participant competency patterns. The consistency between PCA and t-SNE visualizations also strengthens the reliability and stability of the clustering outcomes for competency profiling and learning analytics applications.

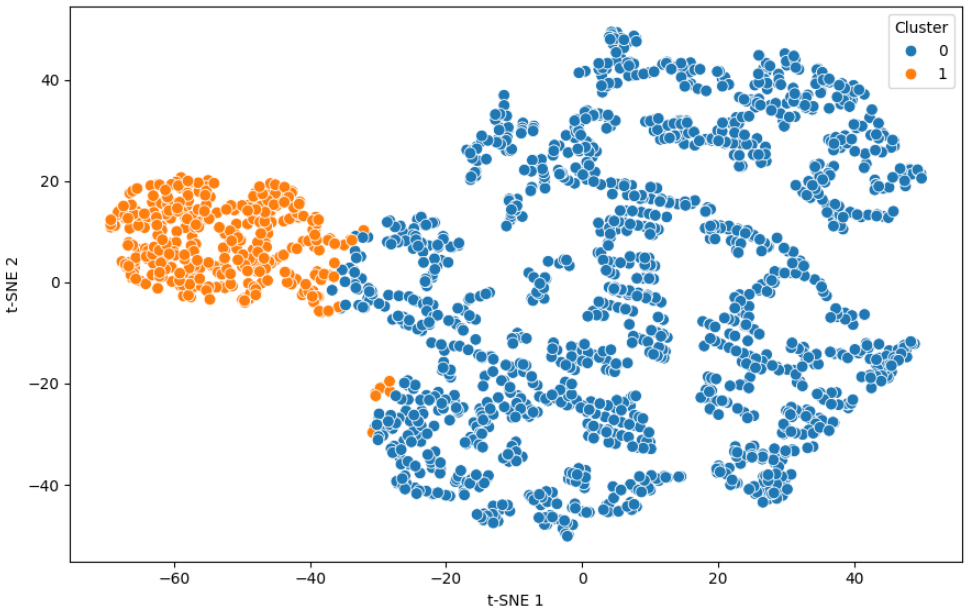


Figure 7 t-SNE Visualization of Candidate Clusters

Total Score Distribution Analysis

Boxplot analysis was conducted to examine the distribution of total SKD scores within each identified cluster, as presented in figure 8. The visualization provides insights into the central tendency, score variability, and distribution spread of participant performance across competency groups. Cluster 0 exhibited a substantially higher median total score compared to Cluster 1, indicating that participants in this cluster generally achieved stronger overall competency performance. In addition, the score distribution within Cluster 0 appeared relatively more stable and concentrated, suggesting greater consistency among participant competency levels. In contrast, Cluster 1 demonstrated lower median scores accompanied by wider score dispersion and greater variability, reflecting more heterogeneous competency characteristics within the lower-performing group. Several outliers were also identified in both clusters, indicating the existence of atypical participant profiles with performance patterns that differ significantly from the majority of participants. These findings further support the effectiveness of the clustering model in distinguishing participant competency groups and identifying variations in performance distribution across clusters.

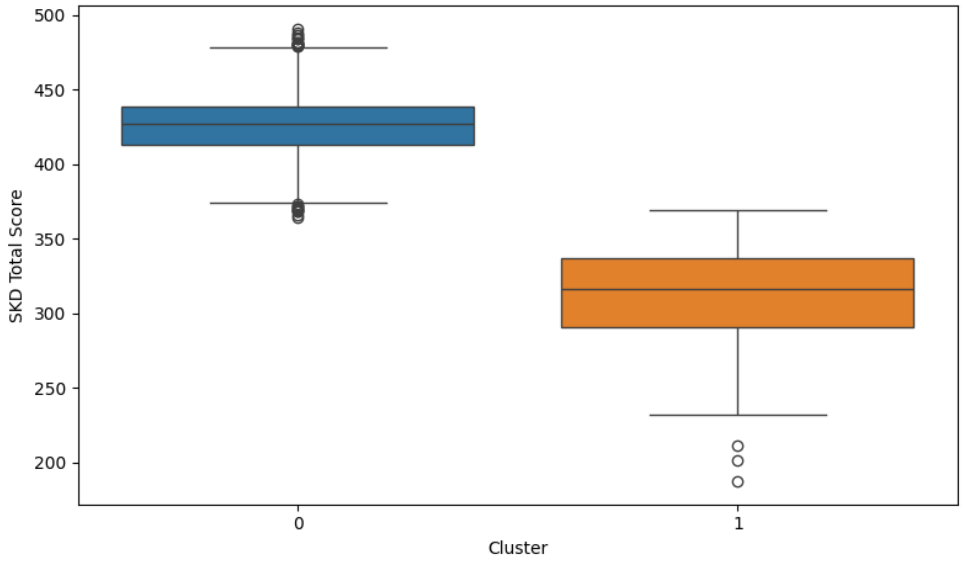


Figure 8 Total Score Distribution by Cluster

Hierarchical Clustering Validation

Hierarchical clustering using Ward linkage was performed to further validate the clustering structure generated by the K-Means algorithm. Ward linkage was selected because it minimizes the variance within clusters and produces compact hierarchical groupings based on Euclidean distance. The resulting dendrogram visualization, presented in figure 9, illustrates the hierarchical relationships among participants and the distance between competency groups. The dendrogram revealed the presence of two dominant hierarchical branches with clear separation at higher linkage distances, indicating that the dataset naturally forms two major competency categories. This hierarchical structure is consistent with the optimal clustering configuration identified through the Elbow Method and Silhouette Score analysis using the K-Means model. The similarity between hierarchical clustering and K-Means results confirms the stability, robustness, and consistency of the identified competency groups, demonstrating that the clustering outcomes are not random but represent meaningful latent structures within the SKD assessment dataset.

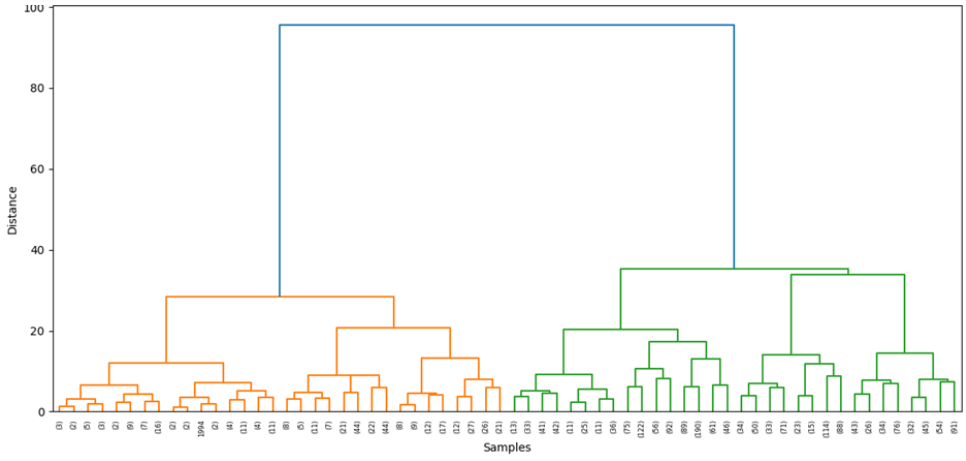


Figure 9 Hierarchical Clustering Dendrogram

Discussion

The findings of this study demonstrate that unsupervised machine learning can effectively identify latent competency structures within SKD assessment data through clustering and learning analytics approaches. The K-Means algorithm successfully separated participants into two distinct competency groups with strong clustering performance, as indicated by the silhouette score, Davies-Bouldin Index, and Calinski-Harabasz Score. The silhouette score of 0.585 indicates good cluster cohesion and separation, suggesting that participants within the same cluster possess similar competency characteristics while remaining sufficiently distinct from participants in other groups. The use of K-Means clustering in educational data mining has been widely reported as effective for grouping learners or participants based on similar performance characteristics and supporting data-driven educational decision-making [21], [22]. The consistency between the Elbow Method, silhouette analysis, PCA visualization, t-SNE representation, and hierarchical clustering further confirms the robustness and stability of the identified cluster structure. Previous studies have also shown that the combination of clustering with dimensionality reduction and visual analytics techniques such as PCA and t-SNE can improve the interpretability of complex educational datasets and reveal meaningful learner or performance profiles [23], [24]. These results indicate that participant competency patterns are not randomly distributed, but instead form naturally distinguishable groups based on similarities in cognitive, civic, and behavioral competencies measured through the SKD assessment. The PCA and t-SNE visualizations also revealed minimal overlap between clusters, indicating that the selected assessment variables contain strong discriminative information capable of representing hidden competency patterns. In the context of learning analytics, these findings demonstrate that visual analytics techniques can provide interpretable representations of participant competency structures and support data driven educational assessment processes.

The competency profiling results revealed substantial differences between clusters across all SKD dimensions, particularly in the TIU component, which emerged as the strongest differentiating factor between participant groups. Cluster 0 consistently achieved higher scores in TWK, TIU, TKP, and total score, indicating stronger overall competency readiness and more balanced performance across assessment dimensions. In contrast, Cluster 1 demonstrated considerably lower competency performance, especially in analytical and cognitive abilities. The large gap observed in TIU scores suggests that intellectual reasoning and analytical capability play a dominant role in determining participant competency profiles within government education selection systems. Furthermore, the heatmap and boxplot analyses showed that participants in Cluster 0 exhibited more stable score distributions, whereas Cluster 1 demonstrated greater variability and heterogeneity in competency characteristics. This finding indicates that lower performing participants tend to possess less consistent competency patterns and may require additional educational intervention or preparatory support. Similar findings in competency-based and performance-based educational contexts suggest that clustering can be used to identify high-performing, moderate-performing, and at-risk participant profiles, thereby supporting targeted intervention and adaptive educational strategies [25]. The DBSCAN analysis also identified several anomalous participant profiles that differed significantly from the majority population, demonstrating the ability of unsupervised learning techniques to detect irregular

competency patterns. Overall, the integration of clustering, visual analytics, and competency interpretation in this study highlights the potential of artificial intelligence-based approaches for supporting intelligent educational assessment, competency mapping, and data driven decision making in educational selection environments.

Conclusion

This study presented an unsupervised machine learning approach for competency profiling in government education selection systems using SKD assessment data. By applying K-Means clustering to TWK, TIU, TKP, and total SKD score variables, the study successfully identified two distinct participant competency groups. The optimal clustering configuration was obtained at two clusters, supported by a silhouette score of 0.585, a Davies-Bouldin Index of 0.683, and a Calinski-Harabasz Score of 2588.239. These evaluation results indicate that the generated clusters have good cohesion, separation, and overall structural quality.

The clustering results showed that Cluster 0 represented participants with higher competency readiness, characterized by consistently higher average scores across all SKD dimensions, while Cluster 1 represented participants with lower competency readiness. Among the assessment variables, TIU emerged as the most discriminative factor, indicating that analytical and cognitive ability played a dominant role in differentiating participant competency profiles. The use of PCA, t-SNE, heatmap visualization, boxplot analysis, and hierarchical clustering further confirmed that the identified competency groups were visually distinguishable and structurally consistent.

Overall, the findings demonstrate that clustering and visual learning analytics can support intelligent competency assessment by revealing hidden participant performance patterns that may not be captured through conventional descriptive analysis alone. This approach can assist educational institutions and government selection systems in conducting data-driven competency mapping, identifying participant readiness levels, and supporting more informed decision-making processes. However, future studies should consider removing redundant variables such as total score from the clustering input, comparing multiple clustering algorithms, validating the results with external assessment outcomes, and incorporating ethical considerations related to fairness, transparency, and data privacy in AI-based educational assessment systems.

Declarations

Author Contributions

Conceptualization: C.J., C.F.; Methodology: C.F.; Software: C.J.; Validation: C.J., C.F.; Formal Analysis: C.J., C.F.; Investigation: C.J.; Resources: C.F.; Data Curation: C.F.; Writing Original Draft Preparation: C.J., C.F.; Writing Review and Editing: C.F., C.J.; Visualization: C.J.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

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

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*, Vancouver, BC, Canada, Apr. 29–May 2, 2012, vol. 2012, no. April, pp. 252–254, doi: 10.1145/2330601.2330661.

[2] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009, doi: 10.5281/zenodo.3554657.

[3] 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.

[4] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "Data mining algorithms to classify students," in *Proceedings of the 1st International Conference on Educational Data Mining (EDM 2008)*, Montreal, QC, Canada, Jun. 20–21, 2008, pp. 8–17. [Online]. Available: https://www.educationaldatamining.org/EDM2008/uploads/proc/1_Romero_3.pdf

[5] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an 'early warning system' for educators: A proof of concept," *Computers & Education*, vol. 54, no. 2, pp. 588–599, 2010, doi: 10.1016/j.compedu.2009.09.008.

[6] G. Siemens, "Learning analytics: The emergence of a discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, 2013, doi: 10.1177/0002764213498851.

[7] D. Ifenthaler and D. Gibson, Eds., *Adoption of Data Analytics in Higher Education Learning and Teaching*. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-47392-1.

[8] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, no. April, pp. 13–49, 2019, doi: 10.1016/j.tele.2019.01.007.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012. [Online]. Available: https://www.sciencedirect.com/book/monograph/9780123814791/data-mining-

concepts-and-techniques

[10] A. Peña-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works,” *Expert Systems with Applications*, vol. 41, no. 4, pt. 1, pp. 1432–1462, 2014, doi: 10.1016/j.eswa.2013.08.042.

[11] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979, doi: 10.2307/2346830.

[12] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Berlin, Germany: Springer, 2001, doi: 10.1007/978-3-642-56927-2.

[13] R. Xu and D. C. Wunsch, “Survey of clustering algorithms,” *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005, doi: 10.1109/TNN.2005.845141.

[14] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: https://jmlr.org/papers/v9/vandermaaten08a.html

[15] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002, doi: 10.1007/b98835.

[16] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J. L. Santos, “Learning analytics dashboard applications,” *American Behavioral Scientist*, vol. 57, no. 10, pp. 1500–1509, 2013, doi: 10.1177/0002764213479363.

[17] W. Greller and H. Drachsler, “Translating learning into numbers: A generic framework for learning analytics,” *Educational Technology & Society*, vol. 15, no. 3, pp. 42–57, 2012. [Online]. Available: https://eric.ed.gov/?id=EJ992502

[18] S. K. Mohamad and Z. Tasir, “Educational data mining: A review,” *Procedia - Social and Behavioral Sciences*, vol. 97, no. November, pp. 320–324, 2013, doi: 10.1016/j.sbspro.2013.10.240.

[19] R. Ferguson, “Learning analytics: Drivers, developments and challenges,” *International Journal of Technology Enhanced Learning*, vol. 4, no. 5/6, pp. 304–317, 2012, doi: 10.1504/IJTEL.2012.051816.

[20] A. Klašnja-Milićević, M. Ivanović, and A. Nanopoulos, “Recommender systems in e-learning environments: A survey of the state-of-the-art and possible extensions,” *Artificial Intelligence Review*, vol. 44, no. 4, pp. 571–604, 2015, doi: 10.1007/s10462-015-9440-z.

[21] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, “K-means algorithm for clustering of learners performance levels using machine learning techniques,” *Revue d'Intelligence Artificielle*, vol. 35, no. 1, pp. 99–104, 2021, doi: 10.18280/ria.350112.

[22] R. Liu, “Data analysis of educational evaluation using k-means clustering method,” *Computational Intelligence and Neuroscience*, vol. 2022, no. July, Art. no. 3762431, 2022, doi: 10.1155/2022/3762431.

[23] E. Priyanto, Berlilana, and I. Tahyudin, “Enhancing clustering performance through benchmarking of dimensionality reduction techniques on educational data,” *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 2, pp. 641–654, Apr. 2025, doi: 10.52436/1.jutif.2025.6.2.4297.

[24] 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.

[25] D. Wahdiana, Ekohariadi, and R. Suhartini, “Clustering cadet training performance using k-means and Ward's method evidence from OTMon maritime monitoring system,” *Journal of Applied Data Sciences*, vol. 6, no. 4, pp. 2741–2757, Dec. 2025, doi: 10.47738/jads.v6i4.958.