ISSN: 2623-1662 (online), 2460-092X (printed)
DOI: https://doi.org/10.19109/jusifo.v11i2.31031
Vol. 11, No. 2 (2025), p.133-140

JUSIFO
JURNAL SISTEM INFORMASI

Research Article [OPEN ACCESS]

Clustering-Based Identification of Student Support Needs in Higher Education Transition

Mochamad Welly Rosadi*, Nenden Siti Fatonah, Gerry Firmansyah, Habibullah Akbar

ABSTRACT

The transition from secondary to higher education represents a critical phase influenced by both academic readiness and socio-economic conditions. This study proposes a clustering-based approach to identify student support needs during this transition by analyzing multidimensional student profiles. Using secondary data from 1,226 senior high school students, three unsupervised clustering algorithms—K-Means, DBSCAN, and BIRCH—were applied to academic performance and socio-economic variables. Cluster quality was assessed using internal validation metrics, including the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. The results indicate that clustering-based methods provide richer insights than traditional rule-based approaches by capturing heterogeneous student profiles and revealing atypical cases. Among the evaluated algorithms, BIRCH demonstrated the most balanced performance in terms of cluster compactness and separation, while K-Means offered stable and interpretable results, and DBSCAN was effective in identifying outliers. Interpreted within the college readiness framework, the identified clusters highlight differentiated student support needs, enabling more targeted and equitable intervention strategies. These findings underscore the potential of educational data mining to support data-driven decision-making in facilitating students' transition to higher education.

Keyword: Clustering analysis, college readiness, educational data mining

Received: August 20, 2025; Revised: December 11, 2025; Accepted: December 27, 2025

Corresponding Author: Mochamad Welly Rosadi, Department of Computer Science, Universitas Esa Unggul, Indonesia, welly.rosadi@student.esaunggul.ac.id
Authors: Nenden Siti Fatonah, Department of Computer Science, Universitas Esa Unggul, Indonesia, nenden.siti@esaunggul.ac.id; Gerry Firmansyah, Department of Computer Science, Universitas Esa Unggul, Indonesia, gerry@esaunggul.ac.id; Habibullah Akbar, Department of Computer Science, Universitas Esa Unggul, Indonesia, habibullah.akbar@esaunggul.ac.id

1. INTRODUCTION

The transition from secondary education to higher education constitutes a pivotal stage in students' educational trajectories, as it determines access to advanced learning opportunities, professional pathways, and long-term social mobility. Research on educational transitions consistently shows that successful progression to higher education is not driven by academic performance alone, but by a combination of cognitive readiness, socio-economic conditions, and institutional support mechanisms (Conley & French, 2014; OECD, 2018). When these dimensions are not adequately addressed, students who formally complete secondary education may still face substantial barriers to entering higher education.

Empirical studies further highlight that transition challenges are particularly pronounced in contexts where socio-economic disparities intersect with limited institutional guidance. Lombard (2020) emphasize that students' transitions to higher education are shaped by both academic preparedness and non-academic factors, including financial resources, family background, and access to targeted support programs. Their findings suggest that interventions focusing solely on academic merit are insufficient, as students with comparable academic profiles may experience divergent transition outcomes depending on their socio-economic and support environments. This reinforces the need for a more holistic understanding of student readiness during the transition phase.

Despite this complexity, many secondary education institutions continue to rely on rule-based or threshold-driven mechanisms to identify students in need of assistance. Common practices include the use of minimum grade requirements or standardized examination scores as primary indicators of readiness for higher education. While administratively efficient, such approaches are inherently reductionist and fail to capture latent student profiles that emerge from the interaction of academic and socio-economic factors (Conley & French, 2014; Lombard, 2020). As a result, students with strong academic potential but limited financial or social support may remain under-identified, while others receive generalized interventions that do not align with their specific needs.

In response to these limitations, Educational Data Mining (EDM) has gained prominence as a data-driven approach for analyzing complex educational datasets (Dutt et al., 2015; Kosztyán et al., 2020; Liu, 2022). EDM techniques enable the exploration of multidimensional student data to uncover hidden patterns and relationships that are not readily observable through manual analysis (Romero & Ventura, 2010). Among these techniques, clustering—an unsupervised learning method—has been widely applied to group students based on shared characteristics without predefined labels (Ester et al., 1996; MacQueen, 1967). In educational research, clustering has been used to analyze academic performance (Mohamed Nafuri et al., 2022), identify behavioral patterns (Mohd Talib et al., 2023), and segment learning profiles (Maylawati et al., 2020).

However, existing studies predominantly focus on higher education populations (Cahapin et al., 2023; Cheng & Shwe, 2019; Hooshyar et al., 2020; Wang, 2022) or employ a single clustering algorithm with limited consideration of socio-economic variables. Comparative analyses that examine different clustering paradigms for identifying student support needs at the secondary–tertiary transition stage remain scarce. Moreover, prior research often prioritizes algorithmic performance over the interpretability of clusters and their relevance for institutional decision-making. Consequently, the potential of clustering to inform differentiated support strategies during the transition to higher education has not been fully realized, particularly in contexts characterized by socio-economic heterogeneity.

To address this gap, this study adopts a clustering-based approach to identify student support profiles during the transition to higher education. Specifically, it conducts a comparative analysis of three clustering algorithms—K-Means, DBSCAN, and BIRCH—representing centroid-based, density-based, and hierarchical paradigms, respectively (Ester et al., 1996; MacQueen, 1967; Zhang et al., 1996). By integrating academic performance indicators with socio-economic attributes, this research aims to generate interpretable student groupings that align with the multifaceted nature of transition challenges identified by Lombard (2020). Through this approach, the study contributes to the educational data mining literature while offering practical insights to support more equitable and targeted interventions for students entering higher education.

2. MATERIALS AND METHODS

2.1 Materials

This study utilized secondary data obtained from the academic and administrative information systems of a public senior high school located in suburban Tangerang, Indonesia. The dataset was compiled from two official institutional sources: the national education database (Dapodik) and the electronic academic reporting system (e-Rapor). These systems are routinely employed by Indonesian schools to document students' academic performance and socio-economic background in a standardized manner.

The final dataset comprised 1,226 student records from grades X to XII who had completed secondary education and were eligible for transition to higher education. Each record included three key variables selected to represent students' academic readiness and socio-economic conditions. Academic readiness was operationalized using the average report card grade, recorded on a numeric scale ranging from 0 to 100. Socio-economic conditions were captured through parental income, recorded as an ordinal variable ranging from 0 (no income) to 6 (monthly income exceeding IDR 20,000,000), and the number of household dependents.

The selection of these variables was theoretically grounded in prior research identifying academic achievement and family socio-economic background as dominant determinants of students' transition to higher education. Other potential variables, such as learning motivation or attendance records, were excluded due to incomplete or inconsistent data across institutional systems. A descriptive summary of the dataset is presented in Table 1, which reports the mean, standard deviation, and range for each variable. As shown in Table 1, students achieved an average report card grade of 82.86 (SD = 3.08), with a mean parental income category of 3.49 and an average household dependency of 2.35.

Table 1. Descriptive statistics of student attributes

Attribute	**Mean**	**Standard Deviation**	**Min**	**Max**	**Description**
Report card grade	82.86	3.08	64.48	91.03	Academic performance (0–100 scale)
Parental income (ordinal 0–6)	3.49	1.45	0	6	Ordinal income category (0 = no income; 6 = > IDR 20,000,000)
Number of dependents	2.35	1.27	0	8	Number of household dependents

2.2 Methods

This study adopted an unsupervised learning approach using clustering techniques to identify groups of students with similar academic and socio-economic profiles. Clustering was selected because it does not require predefined class labels and is therefore suitable for exploratory analysis of heterogeneous student populations. The overall analytical procedure followed the research workflow illustrated in Figure 1, which consists of data preparation, clustering, and evaluation stages.

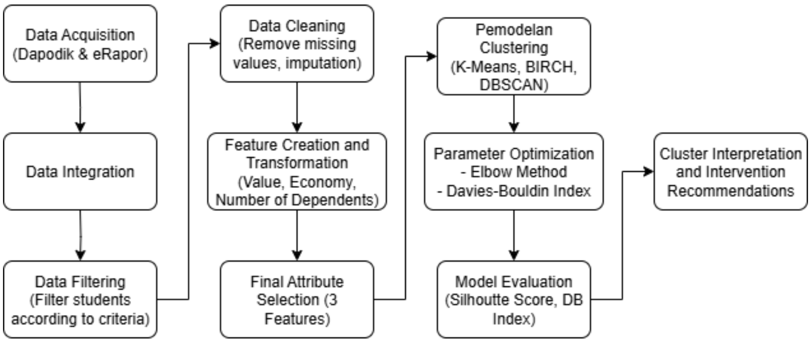
Data Acquisition (Dapodik & eRapor)
Data Integration
Data Filtering (Filter students according to criteria)
Data Cleaning (Remove missing values, imputation)
Feature Creation and Transformation (Value, Economy, Number of Dependents)
Final Attribute Selection (3 Features)
Pemodelan Clustering (K-Means, BIRCH, DBSCAN)
Parameter Optimization - Elbow Method - Davies-Bouldin Index
Model Evaluation (Silhoutte Score, DB Index)
Cluster Interpretation and Intervention Recommendations

Figure 1. Research workflow

Three clustering algorithms were employed to represent different clustering paradigms: K-Means, DBSCAN, and BIRCH. K-Means is a centroid-based algorithm that partitions data into k clusters by minimizing within-cluster variance (MacQueen, 1967). Although computationally efficient and interpretable, K-Means is sensitive to outliers and assumes spherical cluster shapes. In this study, the optimal number of clusters was determined using the elbow method and silhouette coefficient.

DBSCAN is a density-based clustering algorithm capable of identifying clusters with arbitrary shapes and detecting noise points that do not belong to any cluster (Ester et al., 1996). Its performance depends on two parameters, epsilon (eps) and min_samples, which define neighborhood density. Parameter tuning was conducted experimentally to assess cluster stability under different density thresholds.

BIRCH is a hierarchical clustering algorithm designed for large datasets, employing a Clustering Feature (CF) tree to incrementally summarize data points (Zhang et al., 1996). BIRCH was included to evaluate whether a hierarchical approach could generate more compact and interpretable clusters in an educational setting characterized by mixed academic and socio-economic attributes.

Prior to clustering, data preprocessing was conducted following standard practices in educational data mining. Student records from Dapodik and e-Rapor were integrated using the National Student

Identification Number (NISN). Missing values in numeric attributes were imputed using the mean, while duplicate records were removed. All variables were normalized using Min–Max scaling to ensure comparability across features. For DBSCAN, additional experiments confirmed that standardization improved distance-based sensitivity.

Clustering quality was evaluated using three internal validation metrics: Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index, which assess cluster cohesion and separation without external labels (Caliñski & Harabasz, 1974; Davies & Bouldin, 1979; Rousseeuw, 1987). These metrics provided a quantitative basis for comparing clustering outcomes across algorithms, complemented by visualization techniques discussed in the subsequent section.

3. RESULTS AND DISCUSSION

This section presents and discusses the results of the clustering analysis conducted to identify student support profiles during the transition to higher education. The discussion is organized into four subsections. First, the performance of the clustering algorithms is evaluated using internal validation metrics. Second, the characteristics of the resulting clusters are interpreted in relation to students' academic and socio-economic conditions. Third, a comparative discussion highlights the added value of clustering compared to rule-based approaches and situates the findings within prior studies. Finally, visualizations are used to support interpretation, followed by a discussion of practical implications and study limitations.

3.1 Clustering Performance Evaluation

The application of three clustering algorithms—K-Means, DBSCAN, and BIRCH—resulted in distinct clustering structures and evaluation outcomes. The performance of each algorithm was assessed using three internal validation metrics: Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. The comparative results are summarized in Table 2.

Table 2. Clustering performance comparison

Algorithm	Silhouette Score	Davies–Bouldin Index	Calinski–Harabasz Index	Notes
K-Means	0.41	0.87	412.5	Stable across runs
DBSCAN	0.36	0.95	365.2	Sensitive to parameter settings
BIRCH	0.45	0.81	438.9	Best overall balance

As shown in Table 2, BIRCH achieved the highest Silhouette Score (0.45) and Calinski–Harabasz Index (438.9), indicating more compact and well-separated clusters compared to the other methods. K-Means produced stable clustering results with moderate evaluation scores, confirming its suitability as a baseline method. In contrast, DBSCAN demonstrated lower overall performance and higher sensitivity to parameter selection, particularly with respect to eps and min_samples, which affected cluster stability across experiments.

3.2 Cluster Characterization and Student Profiles

Beyond numerical performance, cluster interpretation provides substantive insights into student characteristics and support needs. Using K-Means as an illustrative baseline, students were partitioned into three primary clusters. The first cluster comprised students with high academic achievement and relatively stable socio-economic conditions, indicating strong readiness for higher education. The second cluster included students with moderate academic performance and average family income, suggesting potential benefit from academic guidance or preparatory support. The third cluster consisted of students with lower academic achievement and greater economic constraints, representing a group at higher risk of not transitioning to higher education.

DBSCAN revealed additional nuances by identifying outlier cases that were not captured by centroid-based clustering. These included students with high academic performance but low family income, as well as students with relatively high income but weak academic outcomes. Such cases are often overlooked in rule-based systems and highlight the value of density-based clustering in uncovering atypical yet policy-relevant student profiles.

Among the three algorithms, BIRCH produced the most balanced and interpretable segmentation. Its hierarchical structure allowed clusters to align more closely with practical categories of institutional support, such as scholarship eligibility, academic tutoring needs, and motivational or counseling interventions. This balance between structural quality and interpretability makes BIRCH particularly suitable for educational decision-making contexts.

3.3 Comparative Discussion and Educational Implications

A comparative analysis of the three clustering approaches demonstrates that clustering provides richer insights than traditional rule-based identification methods, which typically rely on fixed grade thresholds. By integrating academic and socio-economic variables, the clustering analysis captures multidimensional student profiles and supports more differentiated intervention strategies.

The superior performance of BIRCH in this study is consistent with prior research indicating that hierarchical clustering methods are effective for educational datasets with mixed attributes (Maylawati et al., 2020; Mohamed Nafuri et al., 2022). Meanwhile, the ability of DBSCAN to detect noise underscores the importance of outlier identification in educational data mining, as atypical student cases often require targeted and non-standard forms of support.

From a practical perspective, the findings suggest clear implications for schools seeking to improve student transitions to higher education. Clusters characterized by strong academic potential but limited economic resources should be prioritized for financial assistance and scholarship programs, while clusters with sufficient economic support but weaker academic performance may benefit more from tutoring, mentoring, or remedial instruction. Such differentiation enables institutions to move beyond one-size-fits-all interventions toward more efficient and equitable support strategies.

3.4 Visualization, Interpretation Support, and Limitations

To support interpretability, clustering results were visualized using three-dimensional scatter plots and a heatmap. As illustrated in Figure 2, the K-Means clustering results show three distinguishable student groups, although partial overlap is observed in the mid-range, reflecting the algorithm's assumption of spherical cluster boundaries. Figure 3 highlights DBSCAN's ability to identify noise points, representing students with exceptional academic–socio-economic combinations.

The BIRCH clustering results, presented in Figure 4, demonstrate more compact and well-separated clusters with minimal overlap, reinforcing the quantitative evaluation outcomes. To further contextualize these results, a heatmap was generated to illustrate the distribution of students' continuation to higher education across K-Means clusters (Figure 5). Cluster 2 exhibited a relatively higher proportion of students continuing to higher education, whereas Clusters 0 and 1 were dominated by students who did not proceed, indicating varying levels of transition risk across clusters.

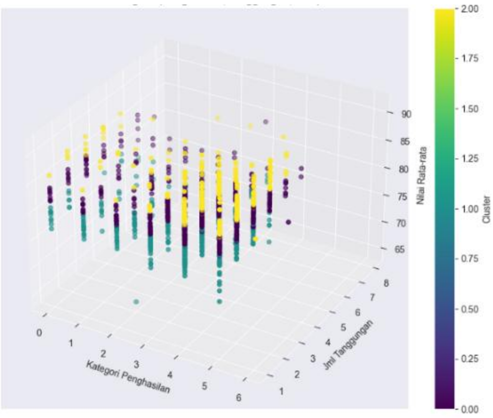

Figure 2. Scatter plot of student clusters using K-Means

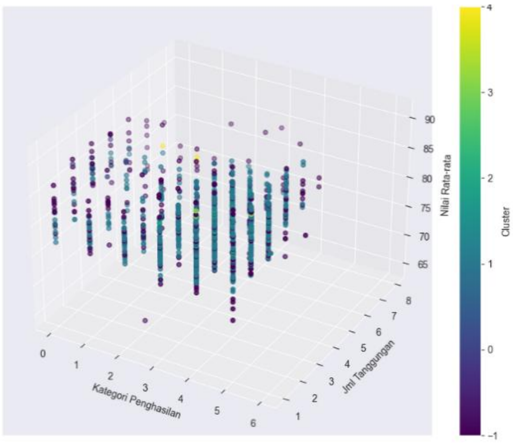

Figure 3. Scatter plot of DBSCAN clustering with noise points identified

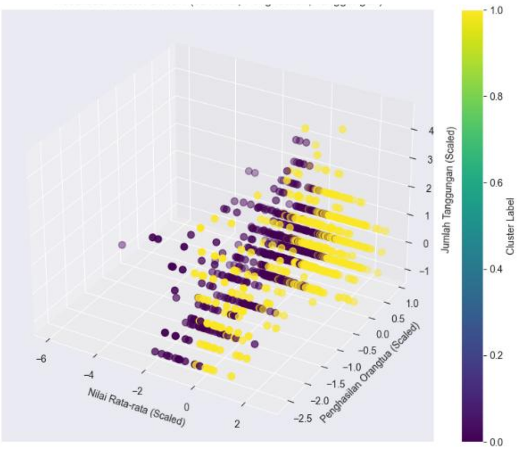

Figure 4. Scatter plot of BIRCH clustering results

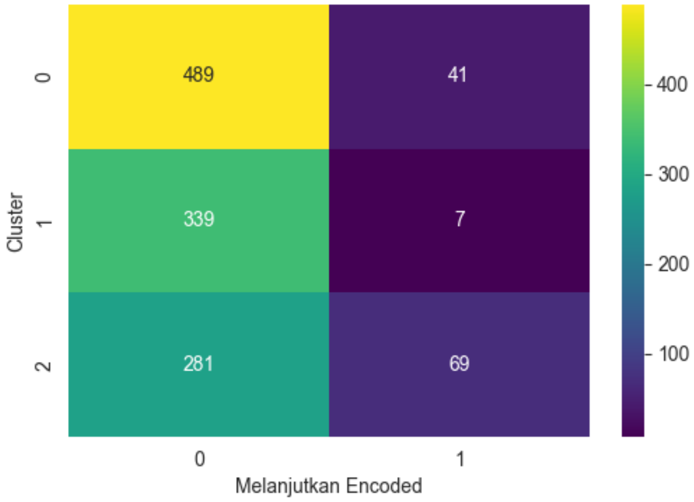

Figure 5. Heatmap of student distribution by cluster

Despite these promising findings, several limitations should be acknowledged. The dataset was derived from a single public high school in suburban Tangerang, which may limit the generalizability of the results to other regions or institutional contexts. Variations in socio-economic composition, institutional resources, and academic environments may influence clustering structures when applied to broader datasets. Future research is encouraged to incorporate multi-school or multi-regional data to enhance external validity and to explore longitudinal clustering approaches that capture changes in student profiles over time.

3.5 Interpretation of Clustering Results in the Context of College Readiness

The clustering results can be coherently interpreted within the framework of college readiness and academic preparedness as articulated by Conley & French (2014). In this framework, readiness for higher education is understood as a multidimensional construct in which academic achievement interacts with contextual conditions that support or constrain students' transition to post-secondary education. Rather than viewing readiness as a single academic threshold, Conley & French emphasizes the importance of both cognitive performance and enabling conditions that facilitate successful educational progression.

Within this perspective, each identified cluster in the present study represents a distinct configuration of academic performance and socio-economic context that shapes students' capacity to continue to higher education. Clusters characterized by higher report card grades and relatively stable family economic conditions reflect stronger overall academic preparedness and more favorable transition contexts.

Conversely, clusters associated with lower academic performance and limited economic resources indicate reduced preparedness and a higher risk of transition discontinuity. These patterns align with Conley & French's argument that academic achievement alone is insufficient to ensure successful transition in the absence of adequate contextual support.

Importantly, the clustering results illustrate how variations in academic and socio-economic profiles translate into differentiated support needs during the transition to higher education. Students with strong academic performance but constrained economic conditions may face barriers primarily related to access and affordability, while students with sufficient economic support but weaker academic outcomes are more likely to require academic reinforcement or structured guidance. By situating these empirically derived student profiles within Conley & French's college readiness framework, this study moves beyond a purely technical clustering analysis and provides a conceptually grounded interpretation of how data-driven methods can support targeted and equitable intervention strategies in secondary education settings.

4. CONCLUSION

This study applied three clustering algorithms—K-Means, DBSCAN, and BIRCH—to analyze students' academic and socio-economic data in order to identify patterns associated with continuation to higher education. The results indicate that clustering-based approaches provide richer insights than traditional rule-based methods by capturing multidimensional student profiles and identifying atypical cases. Among the evaluated algorithms, BIRCH demonstrated the most balanced performance in terms of cluster quality, while K-Means offered stable and interpretable results, and DBSCAN proved particularly valuable for detecting outliers.

From both theoretical and practical perspectives, the findings underscore the potential of clustering methods in educational data mining to support data-driven decision-making. The integration of academic and socio-economic variables enables the identification of differentiated student support needs, facilitating targeted interventions such as financial assistance for academically capable students with limited resources and academic reinforcement for those with adequate resources but weaker performance. Overall, this study highlights the importance of employing multiple clustering techniques and validation metrics to enhance the robustness and interpretability of student segmentation, while future research may extend this approach through additional attributes or longitudinal analysis to further refine transition support strategies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Cahapin, E. L., Malabag, B. A., Santiago, C. S., Reyes, J. L., Legaspi, G. S., & Adrales, K. L. (2023). Clustering of students admission data using k-means, hierarchical, and dbscan algorithms. *Bulletin of Electrical Engineering and Informatics*, *12*(6), 3647–3656. https://doi.org/10.11591/EEI.V12I6.4849

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Cheng, W., & Shwe, T. (2019). Clustering analysis of student learning outcomes based on education data. *Proceedings - Frontiers in Education Conference, FIE, 2019-October*. https://doi.org/10.1109/FIE43999.2019.9028400

Conley, D. T., & French, E. M. (2014). Student ownership of learning as a key component of college readiness. *American Behavioral Scientist*, *58*(8), 1018–1034. https://doi.org/10.1177/0002764213515232

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, *5*(2), 112–116.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Hooshyar, D., Yang, Y., Pedaste, M., & Huang, Y. M. (2020). Clustering algorithms in an educational context: an automatic comparative approach. *IEEE Access*, *8*, 146994–147014. https://doi.org/10.1109/ACCESS.2020.3014948

Kosztyán, Z. T., Orbán-Mihálykó, Mihálykó, C., Csányi, V. V., & Telcs, A. (2020). Analyzing and clustering students' application preferences in higher education. *Journal of Applied Statistics*, *47*(16), 2961–2983. https://doi.org/10.1080/02664763.2019.1709052

Liu, R. (2022). Data analysis of educational evaluation using k-means clustering method. *Computational Intelligence and Neuroscience*, *2022*(1), 3762431. https://doi.org/10.1155/2022/3762431

Lombard, P. (2020). Factors that influence transition from high school to higher education: a case of the juniortukkie programme. *African Journal of Career Development*, *2*(1). https://doi.org/10.4102/AJCD.V2I1.5

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

Maylawati, D. S., Priatna, T., Sugilar, H., & Ramdhani, M. A. (2020). Data science for digital culture improvement in higher education using k-means clustering and text analytics. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(5), 4569–4580. https://doi.org/10.11591/ijece.v10i5.pp4569-4580

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences*, *12*(19), 9467. https://doi.org/10.3390/APP12199467

Mohd Talib, N. I., Abd Majid, N. A., & Sahran, S. (2023). Identification of student behavioral patterns in higher education using k-means clustering and support vector machine. *Applied Sciences*, *13*(5), 3267. https://doi.org/10.3390/APP13053267

OECD. (2018). Education at a glance 2018: oecd indicators. In *Education at a Glance* (Vol. 2018). OECD Publishing, Paris. https://doi.org/10.1787/EAG-2018-EN

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Wang, Z. (2022). Higher education management and student achievement assessment method based on clustering algorithm. *Computational Intelligence and Neuroscience*, *2022*(1), 4703975. https://doi.org/10.1155/2022/4703975

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). Birch: an efficient data clustering method for very large databases. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *25*(2), 103–114. https://doi.org/10.1145/235968.233324