Jatilima : Jurnal Multimedia Dan Teknologi Informasi
Volume 07, Number 06, 2026, https://doi.org/10.54209/jatilima.v7i06.2282
E.SSN : 2721-1800 (Online)
https://journal.cattleyadf.org/index.php/jatilima/index

A K-Prototypes-Based Approach for Modelling Student Segmentation Based on Learning Strategies to Support Academic Decision-Making

Nurul Ain Farhana[1], Putri Maulidina Fadilah[2], Putri Harliana[3], Suwanto[4]
[1,2]Department of Statistics, [3]Departments of Computer Science, [4]Department of Mathematics Education, Universitas Negeri Medan, Jl. Willem Iskandar, Medan, Indonesia

Article Info

Keywords:
student segmentation; learning strategies; K-Prototypes; clustering; educational data mining

ABSTRACT

This study models student segmentation based on learning strategies using the K-Prototypes clustering algorithm on mixed-type data, including categorical variables (gender, major) and numerical variables such as GPA, learning habits, motivation, learning environment, health, social support, academic involvement, and achievement. The analysis involved data preprocessing, exploratory analysis, and clustering. The optimal number of clusters was determined using the Elbow and Silhouette methods, both indicating four clusters. The results identify four distinct student groups: Cluster 3 represents highly motivated and high-achieving students; Cluster 1 includes students with good performance and supportive learning conditions; Cluster 4 reflects moderate characteristics; and Cluster 2 consists of students with lower performance and weaker learning strategies. t-SNE visualization confirms a reasonably clear cluster distribution despite minor overlap. These findings demonstrate that K-Prototypes effectively handles mixed-type educational data and provides valuable insights for data-driven academic decision-making and targeted learning strategies.

Corresponding Author:
Nurul Ain Farhana
Universitas Negeri Medan
Jl. Willem Iskandar, Pasar V Medan Estate, Sumatera Utara
Email : nurulainfarhana@unimed.ac.id

INTRODUCTION

The development of higher education in the digital era requires more adaptive and data-driven learning approaches. Students, as the primary subjects in the learning process, exhibit diverse characteristics in terms of motivation, learning habits, academic engagement, and learning environments, all of which influence their academic performance. This diversity poses a challenge for educational institutions in designing effective and targeted learning strategies. Student learning strategies play a crucial role in determining academic success. Previous studies have shown that differences in learning strategies, such as levels of engagement, time management, and motivation, can lead to significant variations in academic performance [1]. Furthermore, in modern learning contexts such as blended learning and e-learning, students' ability to regulate their own learning strategies has become increasingly important in achieving successful learning outcomes [2], [3].

However, conventional approaches in identifying student characteristics tend to be generalized and often fail to capture the complexity of multidimensional data. Therefore, a data mining-based approach is required to process large datasets and uncover hidden patterns in student data. One of the widely used techniques in this context is clustering analysis, which aims to group objects based on the similarity of their characteristics [4].

In recent years, the application of clustering techniques in education has grown significantly, particularly in identifying learning behaviour patterns, levels of engagement, and student segmentation to support personalized learning. This approach has been proven effective in generating distinct student profiles, such as highly engaged, moderately engaged, and less engaged students, which are correlated with their academic performance [5]. Additionally, clustering enables the development of more adaptive and need-based learning strategies tailored to different groups of students [6].

Despite these advancements, most previous studies have relied on clustering methods limited to numerical data, such as K-Means, which are less effective in handling mixed-type data. In practice, student data typically consist of both numerical and categorical variables, such as gender, major, and behavioural indicators. Therefore, a method capable of handling both types of data simultaneously is required.

The K-Prototypes method is an extension of clustering algorithms designed to handle mixed data types by combining the strengths of K-Means and K-Modes [7]. This method enables a more representative clustering process in educational contexts, particularly in identifying complex and multidimensional student learning strategies.

Based on the above discussion, this study aims to model student segmentation based on learning strategies using the K-Prototypes method. The results of this study are expected to contribute to supporting more accurate and data-driven academic decision-making, particularly in designing learning strategies that align with student characteristics.

METHODS

Educational Data Mining in Higher Education

Educational Data Mining (EDM) has become an essential approach in analysing large-scale educational data to enhance learning outcomes and support institutional decision-making. EDM focuses on extracting meaningful patterns from student-related data, including academic performance, engagement, and behavioural attributes [8]. With the rapid growth of digital learning environments, such as Learning Management Systems (LMS), a large amount of educational data is generated. This has encouraged the application of data mining techniques to better understand student learning behaviours and improve the effectiveness of teaching strategies [9].

Clustering Techniques in Educational Contexts

Clustering is a widely used unsupervised learning technique that aims to group data objects based on similarity. In educational contexts, clustering has been applied to classify students according to their academic performance, engagement levels, and learning strategies [10].

Previous studies have demonstrated that clustering can effectively identify groups such as high-performing students, moderate learners, and at-risk students. These classifications

provide valuable insights for designing personalized learning strategies and targeted academic interventions [11], [12]. Furthermore, clustering techniques support institutions in improving academic decision-making processes [13], [14].

Limitations of Conventional Clustering Methods

Despite the effectiveness of clustering methods such as K-Means, these techniques are primarily designed for numerical data and are not suitable for datasets containing mixed data types. In educational datasets, variables often consist of both numerical and categorical data, making conventional methods less effective [15].

This limitation can lead to suboptimal clustering results and reduced interpretability, especially when dealing with complex and multidimensional student data. Therefore, more advanced clustering methods that can handle mixed data types are required.

Determination of Optimal Number of Clusters

Elbow Method

The Elbow Method determines the optimal number of clusters by analysing the total within-cluster variation. The within-cluster sum of squares (WCSS) is calculated as:

$$WCSS = \sum_{j=1}^{k} \sum_{i \in C_j} \left\| x_i - \mu_j \right\|^2 \quad (1)$$

Where, k = the number of clusters; C_j = the set of data points belonging to the j-th cluster; x_i =the i-th data point in cluster C_j; μ_j = the centroid (mean) of cluster C_j; and $\left\| x_i - \mu_j \right\|^2$ = the squared Euclidean distance between a data point and its cluster centroid. The optimal number of clusters is identified at the point where the reduction in variation becomes less significant, forming an "elbow" pattern [16].

Silhouette Method

The Silhouette method evaluates clustering quality by measuring how similar an object is to its own cluster compared to other clusters. The silhouette coefficient is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \quad (2)$$

Where, $s(i)$ = silhouette value for data point i; $a(i)$ = average distance to points in the same cluster; and $b(i)$ = average distance to the nearest other cluster. A higher silhouette value indicates better clustering quality [17].

K-Prototypes Clustering Algorithm

The K-Prototypes algorithm is an extension of clustering techniques that integrates K-Means and K-Modes to handle mixed data types. This method minimizes a combined dissimilarity function for numerical and categorical variables, allowing simultaneous processing of heterogeneous data [7].

The objective function of the K-Prototypes algorithm is defined as:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \partial_{ij} \left(\sum_{l \in num} \left(x_{il} - z_{jl} \right)^2 + \lambda \sum_{l \in cat} \partial\left(x_{il}, z_{jl} \right) \right) \quad (3)$$

Where: n = number of observations; k = number of clusters; ∂_{ij} = membership indicator; λ = weighting parameter for categorical variables.

This formulation allows simultaneous processing of numerical and categorical attributes, making K-Prototypes highly suitable for educational datasets.

Student Segmentation for Academic Decision-Making

Student segmentation based on clustering results provides valuable insights for academic decision-making. By identifying groups of students with similar learning characteristics, institutions can design adaptive learning strategies tailored to each cluster.

Previous studies have shown that clustering-based segmentation can effectively identify at-risk students, improve academic performance, and support data-driven policy decisions in higher education [13].

RESULTS AND DISCUSSION

Data Description

The dataset used in this study consists of student data from the Faculty of Mathematics and Natural Sciences (FMIPA). The data include nine variables representing student characteristics related to learning strategies. These variables consist of both categorical and numerical types, making them suitable for analysis using the K-Prototypes algorithm.

The variables used in this study, along with their types and descriptions, are presented in Table 1.

Tabel 1. Description of Research Variables

Variable Code	Variable Name
X1	Gender
X2	Major
X3	Grade Poin Average (GPA)
X4	Learning Habits and Time Management
X5	Academic Motivation
X6	Learning Environment
X7	Health and Social Support
X8	Academic and Organizational Involvement
X9	Academic Achievement

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the distribution and characteristics of the dataset. The analysis shows that the numerical variables exhibit varying ranges and distributions, indicating differences in student learning behaviours.

In addition, initial observations suggest that the dataset contains heterogeneous patterns, which supports the use of clustering techniques to identify underlying group structures. No severe anomalies were observed; however, variations across variables indicate the need for careful preprocessing.

Data Preprocessing

Before performing clustering, data preprocessing was conducted to ensure data quality and compatibility with the K-Prototypes algorithm. Categorical variables (X1 and X2) were transformed into factor types, while numerical variables (X3–X9) were converted into numeric format. This step ensures that each variable is processed appropriately during clustering. Furthermore, data cleaning was performed to handle potential inconsistencies and missing values. The preprocessing stage is essential to improve clustering accuracy and avoid bias in the results.

Determination of Optimal Number of Clusters

The determination of the optimal number of clusters in this study was carried out using the Elbow and Silhouette methods. The results of both methods are presented in Figure 1 and Figure 2, respectively.

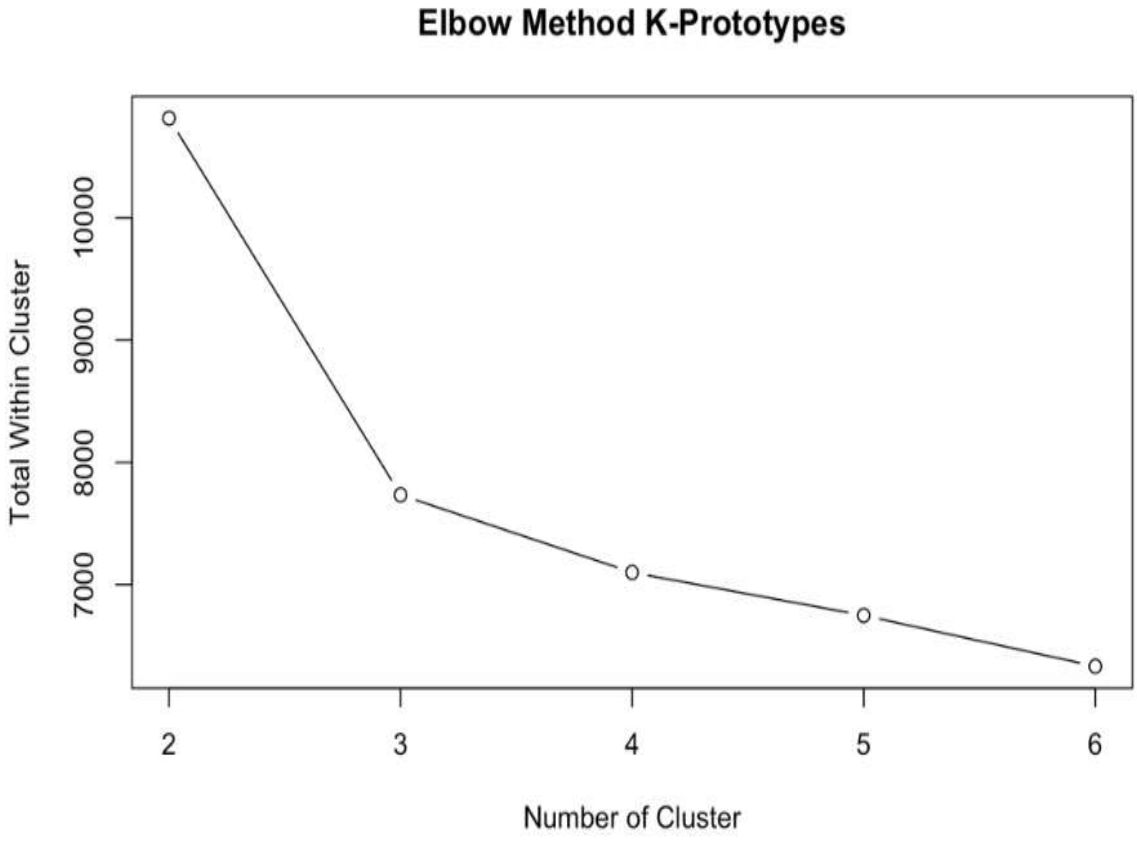

Figure 1. Elbow Method

The Elbow method, as shown in Figure 1, illustrates the relationship between the number of clusters and the total within-cluster sum of squares (WCSS). It can be observed that the WCSS value decreases significantly as the number of clusters increases from $k = 2$ to $k = 4$. However, after $k = 4$, the rate of decrease becomes less pronounced, forming an "elbow" pattern. This indicates that adding more clusters beyond this point does not significantly improve cluster compactness.

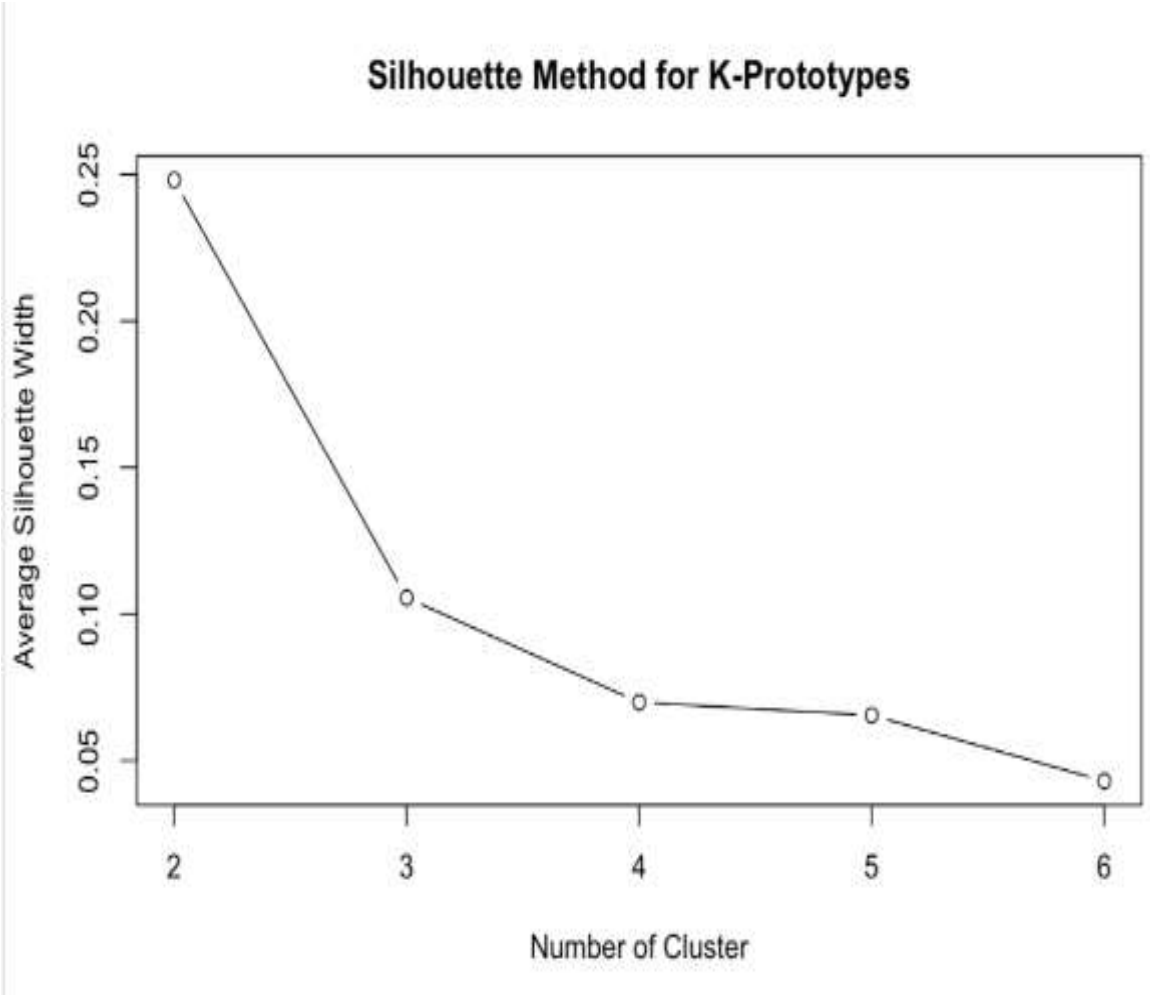

Figure 2. Silhouette Method

Meanwhile, the Silhouette method, presented in Figure 2, evaluates the quality of clustering by measuring the average silhouette coefficient for each value of k. The results show that the silhouette values tend to stabilize and reach a relatively optimal level at $k = 4$. This suggests that the clustering structure at $k = 4$ provides a good balance between intra-cluster cohesion and inter-cluster separation. Based on the results of both the Elbow and Silhouette methods, it can be concluded that the optimal number of clusters for this study is four clusters. This selection ensures that the clustering results are both compact and well-separated, making them suitable for further analysis.

Visualization of Clustering Results

To further interpret the clustering results, both visualization and cluster profile analysis were conducted. The visualization using the t-Distributed Stochastic Neighbor Embedding (t-SNE) method is presented in Figure 3, while the profile of cluster characteristics is shown in Figure 4.

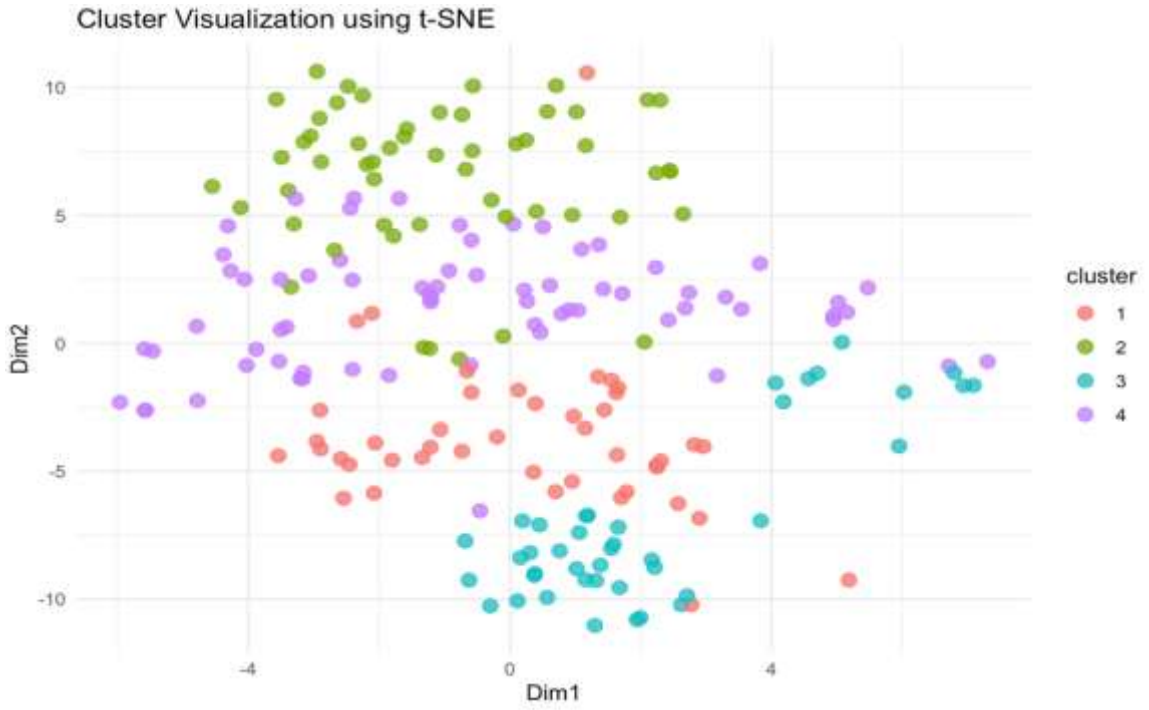

Figure 3. Cluster Visualization

The t-SNE visualization (Figure 3) illustrates the distribution of students across four clusters in a two-dimensional space. Each point represents an individual student, and different colours indicate cluster membership. It can be observed that several clusters, particularly Cluster 2 and Cluster 3, form relatively compact and distinguishable groups. However, a certain degree of overlap is still present, especially between Cluster 1 and Cluster 4, indicating that some students share similar characteristics across clusters. This reflects the inherent complexity of student learning behaviours and suggests that the clustering results capture realistic patterns in the data.

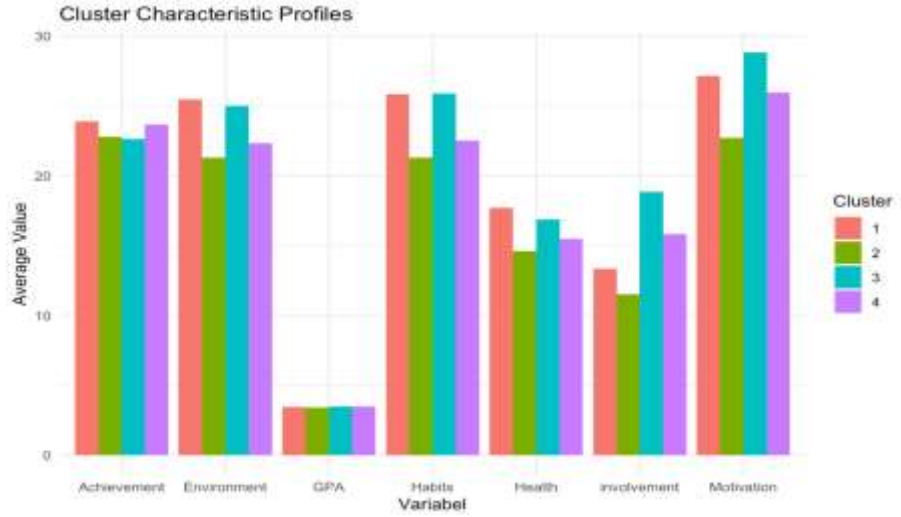

Figure 4. Characteristic Profile of Each Cluster

In addition to visualization, the cluster profiles shown in Figure 4 provide a clearer understanding of the characteristics of each cluster based on the average values of numerical

variables. The variables analysed include GPA, learning habits, health and social support, academic involvement, learning environment, motivation, and academic achievement.

Cluster 3 consistently exhibits the highest values across most variables, particularly in motivation, learning habits, academic involvement, and achievement. This indicates that students in this cluster have strong learning strategies, high engagement, and excellent academic performance. Therefore, Cluster 3 can be identified as the group of highly motivated and high-achieving students.

Cluster 1 shows relatively high values in learning habits, environment, and motivation, although slightly lower than Cluster 3. Students in this cluster demonstrate good academic performance supported by favourable learning conditions, but their level of involvement is comparatively moderate.

Cluster 4 presents moderate values across all variables, indicating a balanced but not outstanding performance. Students in this group have stable learning characteristics and show potential for improvement, particularly in terms of motivation and engagement.

On the other hand, Cluster 2 consistently records the lowest values across most variables, including learning habits, health, involvement, and motivation. This suggests that students in this cluster have weaker learning strategies and lower academic performance, making them a group that requires more attention and targeted academic support. Overall, the combination of t-SNE visualization and cluster profile analysis provides a comprehensive understanding of student segmentation. The visualization highlights the distribution and separation of clusters, while the profile analysis explains the underlying characteristics of each group. These findings reinforce the effectiveness of the K-Prototypes algorithm in identifying meaningful patterns within mixed-type educational data.

CONCLUSION

This study aims to model student segmentation based on learning strategies using the K-Prototypes clustering algorithm on mixed-type data consisting of categorical and numerical variables. The variables analyzed include gender, major, GPA, learning habits, motivation, learning environment, health and social support, academic involvement, and academic achievement. The results of the analysis indicate that the optimal number of clusters is four, as determined using the Elbow and Silhouette methods. The clustering process using the K-Prototypes algorithm successfully grouped students into four distinct clusters with different learning characteristics. Further analysis of cluster profiles reveals meaningful patterns among students. Cluster 3 represents highly motivated and high-achieving students with strong engagement and effective learning strategies. Cluster 1 consists of students with good academic performance supported by favorable learning conditions, although their level of involvement is moderate. Cluster 4 includes students with balanced but average characteristics, indicating potential for improvement. Meanwhile, Cluster 2 represents students with lower performance and weaker learning strategies, requiring more attention and academic support.

The visualization using the t-SNE method supports the clustering results by showing a reasonably clear distribution of clusters, although some overlap exists due to the complexity of student characteristics. This confirms that the clustering results reflect real-world patterns

in educational data. Overall, this study demonstrates that the K-Prototypes algorithm is effective in handling mixed-type data and identifying meaningful student segments. The findings provide valuable insights for academic decision-making, particularly in designing targeted learning strategies, improving student performance, and supporting data-driven educational policies.

ACKNOWLEDGEMENT

The authors are grateful to the Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Negeri Medan for providing the financial resources necessary to conduct this research through their internal grant scheme.

REFERENCE

[1] E. M. Campeanu, I. A. Boitan, and D. G. Anghel, “Student engagement and academic performance in pandemic-driven online teaching: An exploratory and machine learning approach,” *Management & Marketing*, vol. 18, no. s1, pp. 315–339, Dec. 2023, doi: 10.2478/mmcks-2023-0017.

[2] J. Broadbent and W. L. Poon, “Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review,” *Internet High. Educ.*, vol. 27, pp. 1–13, Oct. 2021, doi: 10.1016/j.iheduc.2015.04.007.

[3] A. Bozkurt *et al.*, “A global outlook to the interruption of education due to COVID-19 Pandemic: Navigating in a time of uncertainty and crisis,” *Asian Journal of Distance Education*, vol. 15, no. 1, 2020, [Online]. Available: http://www.asianjde.org

[4] C. C. Aggarwal, *Data Mining: The Textbook*, 3rd ed., vol. 1. 2021.

[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, May 2020, doi: 10.1002/widm.1355.

[6] S. A. Aljawarneh, “Reviewing and exploring innovative ubiquitous learning tools in higher education,” *J. Comput. High. Educ.*, vol. 32, no. 1, pp. 57–73, Apr. 2020, doi: 10.1007/s12528-019-09207-0.

[7] M. Jannadi and K. Ben Driss, “Composed clustering of non-relational data with mixed types using K-modes and K-prototypes algorithms,” in *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, New York, NY, USA: ACM, Oct. 2020, pp. 1–8. doi: 10.1145/3423603.3424050.

[8] D. A. Dewi, “Sustainable Educational Data Mining Studies: Identifying Key Factors and Techniques for Predicting Student Academic Performance,” *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1325–1342, Sep. 2024, doi: 10.47738/jads.v5i3.347.

[9] S. N. Safitri, Haryono Setiadi, and E. Suryani, “Educational Data Mining Using Cluster Analysis Methods and Decision Trees based on Log Mining,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 3, pp. 448–456, Jul. 2022, doi: 10.29207/resti.v6i3.3935.

[10] N. A. Ramadhani and R. D. Hardianti, “Building Students’ Creative Thinking Skills in Science Learning : A Systematic Review of STEM-Based Approaches,” *Jurnal Paedagogy*, vol. 12, no. 3, p. 831, Jul. 2025, doi: 10.33394/jp.v12i3.15964.

[11] A. Abu, “Educational Data Mining & Students’ Performance Prediction,” *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, 2021, doi: 10.14569/IJACSA.2016.070531.
[12] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, “Analyzing and Predicting Students’ Performance by Means of Machine Learning: A Review,” *Applied Sciences*, vol. 10, no. 3, p. 1042, Feb. 2020, doi: 10.3390/app10031042.
[13] 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.
[14] M. Apte and A. Bhave-Gudipudi, “Cooperative Learning techniques to bridge gaps in academia and corporate,” *Procedia Comput. Sci.*, vol. 172, pp. 289–295, 2020, doi: 10.1016/j.procs.2020.05.046.
[15] E. Crisol-Moya, M. J. Caurcel-Cara, P. Peregrina-Nievas, and C. del P. Gallardo-Montes, “Future Mathematics Teachers’ Perceptions towards Inclusion in Secondary Education: University of Granada,” *Educ. Sci. (Basel).*, vol. 13, no. 3, p. 245, Feb. 2023, doi: 10.3390/educsci13030245.
[16] L. Chen, “Innovative Application of Data Mining Technology in College Information System Based on Informatized Teaching Environment,” *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Jan. 2024, doi: 10.2478/amns-2024-1613.
[17] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, “An extensive comparative study of cluster validity indices,” *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2022, doi: 10.1016/j.patcog.2012.07.021.