E-ISSN: 3123-2507
DOI: https://doi.org/10.64803/cessmuds.v1.159 | Page: 762-766
Farhana, Fadilah, Harliana, and Suwanto

Academica Performance Pattern Identification of FMIPA Medan State University Students Using the K-Prototype Clustering Algorithm

Nurul Ain Farhana[1*], Putri Maulidina Fadilah[2], Putri Harliana[3], Suwanto[4]

[1,2,3,4]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Medan State University, Jl. William Iskandar Ps. V, 20221, North Sumatera, Indonesia
E-mail: [1*]nurulainfarhana@unimed.ac.id, [2]putrimaulidina@unimed.ac.id, [3]harliana@unimed.ac.id, [4]suwanto@unimed.ac.id
*E-mail Corresponding Author: nurulainfarhana@unimed.ac.id

Abstract

Abstract: This study aims to identify patterns in academic performance among students at the Faculty of Mathematics and Natural Sciences (FMIPA), Medan State University (UNIMED) using the K-Prototype clustering algorithm. The method was chosen due to its capability to handle mixed data types, combining both numerical and categorical variables. Data were collected through a structured questionnaire, and the sample size was determined using Slovin's formula, resulting in 373 student records for analysis. The clustering process revealed three distinct groups: students with stable academic records but low organizational engagement; highly motivated new students; and independent senior students active in extracurricular activities. Each cluster reflects unique combinations of GPA, credit load, learning behavior, and involvement. The findings provide a data-driven foundation for designing targeted academic advising and student support strategies.

Keywords: *Algorithm, Cluster, K-Prototype, Slovin*

I. INTRODUCTION

Academic achievement serves as a fundamental indicator of student development and institutional quality in higher education. As global competition among universities intensifies, there is increasing pressure to monitor and improve students' academic performance through evidence-based approaches [1], [2]. Universities must not only track academic records but also understand the underlying behavioral and psychological dimensions that shape student outcomes [3].

Numerous studies have highlighted the complexity of academic success, which depends on a combination of cognitive abilities, motivation, study habits, stress management, time allocation, learning resources, and support systems [4], [5]. These variables do not operate in isolation; they interact and influence students in ways that are often difficult to observe using traditional analytic tools.

The field of educational data mining has emerged as a powerful approach for uncovering latent patterns within student data [6]. Clustering, a type of unsupervised machine learning, is widely applied to group students based on shared attributes, facilitating the design of more tailored interventions [7], [8]. However, most popular clustering algorithms like K-Means are restricted to numerical data only and are inadequate when categorical data such as gender, year of enrollment, or academic program is involved [9].

To address these limitations, the K-Prototypes clustering algorithm was developed. It integrates the K-Means approach for numeric features and the K-Modes algorithm for categorical variables, allowing it to handle mixed-type datasets effectively [10]. In educational contexts, this is especially relevant as student data typically include a blend of quantitative (e.g., GPA, credit load) and qualitative (e.g., gender, program of study, motivation levels) features.

This paper focuses on applying the K-Prototypes algorithm to identify academic performance patterns among students of the Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Negeri Medan (UNIMED). A structured dataset was constructed based on survey responses from a representative sample of students. The dataset incorporates not only academic metrics but also various behavioral and motivational indicators.

The outcomes of this clustering analysis are expected to provide insight into how students can be segmented according to their academic trajectories and behavioral tendencies. Ultimately, the findings may contribute to the formulation of adaptive strategies in academic advising, student support programs, and curriculum planning, helping institutions shift from a one-size-fits-all approach to a more personalized model of student success [11].

II. RESEARCH METHODOLOGY

This study employs a quantitative approach using the K-Prototypes clustering algorithm to identify patterns in academic performance among undergraduate students at the Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Negeri Medan. The research framework consists of five main stages: sample determination, data collection, data preprocessing, clustering, and interpretation.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

A. Sample Size and Respondents

The respondents of this study are undergraduate students from the Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Negeri Medan. The total population includes students from all departments and entry cohorts. To determine the appropriate number of respondents, Slovin's formula is utilized to calculate the minimum required sample size [12]:

$$n = \frac{N}{1 + N \cdot e^2} \quad (1)$$

where n is the required sample size, N is the total population, and e is the margin of error (typically set at 0.05).

This formula allows for determining a representative sample when the population size is known, but the population variance is unknown. It ensures that the sample reflects the diversity of the population within an acceptable margin of error.

Respondents were selected using a proportional stratified random sampling technique to ensure balanced representation from various academic programs and entry years. All participants voluntarily completed the questionnaire distributed through official institutional communication channels.

B. Data Preprocessing

The dataset underwent several preprocessing steps to ensure its readiness for clustering analysis using the K-Prototypes algorithm. First, invalid or incomplete responses were removed during data cleaning. Missing values in the Likert-scale items were handled using mean imputation. Categorical variables such as gender, program of study, and year of entry were encoded as factors, while numerical variables including GPA, total credits, and average credits per semester were normalized using Min-Max scaling. Behavioral indicators from Likert-scale responses were also normalized and treated as numerical inputs. The final dataset consisted of 373 complete records with mixed-type features ready for clustering.

C. K-Prototype Clustering

This study employed the K-Prototypes algorithm to cluster students based on mixed-type attributes, including categorical variables (e.g., gender, academic program, year of entry) and numerical variables (e.g., GPA, total credits, behavioral scores). K-Prototypes is an extension of the K-Means algorithm that handles both numerical and categorical data by combining Euclidean distance for numerical features and matching dissimilarity for categorical ones. The objective function minimized in K-Prototypes is [13]:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \delta(z_i, c_j) \quad (2)$$

where $\delta(z_i, c_j)$ is the dissimilarity between data point z_i and prototype c_j, calculated as [16]:

$$\delta(z_i, c_j) = \sum_{l \in num} (z_{il} - c_{jl})^2 + \lambda \sum_{l \in cat} \| (z_{il} \neq c_{jl}) \quad (3)$$

The implementation was carried out in R using the kproto() function from the **klaR** package. The optimal number of clusters k was determined using the elbow method, where the Total Within-Cluster Sum of Squares (TWSS) was plotted against varying $k = 1,2,3, \ldots, n$ values to identify the point where adding more clusters yields diminishing returns. This method effectively grouped students into clusters with similar academic achievements and learning behaviors, allowing for a nuanced analysis of performance patterns.

D. Cluster Evaluation

To assess the quality of the clustering results, the Total Within-Cluster Sum of Squares (TWSS) was used as the main evaluation metric. TWSS measures the compactness of clusters by calculating the sum of squared distances between each data point and its respective cluster centroid or prototype [14], [15]:

$$TWSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (4)$$

Nnnn where C_j the set of points in cluster j, μ_j is the prototype of cluster j and $\|x_i - \mu_j\|$ is Euclidean distance for numerical attributes, with matching dissimilarity for categorical ones.

The Elbow Method was applied by plotting TWSS against different values of k (number of clusters). The optimal number of clusters was determined at the "elbow point" where a significant decrease in TWSS begins to level off indicating a balance between underfitting and overfitting.

In addition to TWSS, each cluster was analysed descriptively by examining the mode of categorical attributes and the mean of numerical attributes, including GPA, SKS, and Likert-based behavioral indicators. This allowed for clear profiling and interpretation of each cluster's characteristics.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

E-ISSN: 3123-2507
DOI: https://doi.org/10.64803/cessmuds.v1.159 | Page: 762-766
Farhana, Fadilah, Harliana, and Suwanto

III. RESULTS AND DISCUSSION

A. Determination of the Optimal Number of Clusters

To ensure an appropriate number of clusters for the K-Prototypes algorithm, the elbow method was employed by computing the Total Within-Cluster Sum of Squares (TWSS) for several values of k. This technique helps identify the point where increasing the number of clusters results in marginal gain, thus avoiding overfitting.

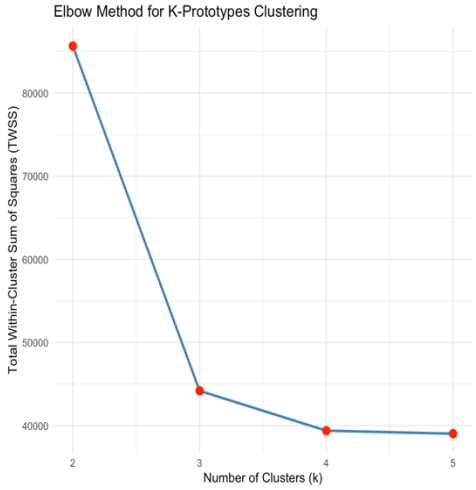


Figure 1. Elbow Plot for Determining the Optimal Number of Clusters Using The K-Prototypes Algorithm

As depicted in Figure 1, the curve shows a sharp drop in TWSS when increasing the number of clusters from k = 2 to k = 3. Beyond this point, the TWSS value begins to plateau, indicating that further increases in the number of clusters would not substantially improve the clustering compactness. Therefore, k = 3 was selected as the optimal number of clusters, balancing model simplicity with accuracy. This choice provides sufficient differentiation among student profiles while avoiding fragmentation of homogeneous groups.

B. Academic Performance Cluster Visualization

The clustering process using k = 3 produced three distinct groups of students based on a combination of numerical and categorical attributes. To gain insight into how the clusters differ in terms of academic achievement, a scatter plot was constructed using GPA and total number of credits (SKS).

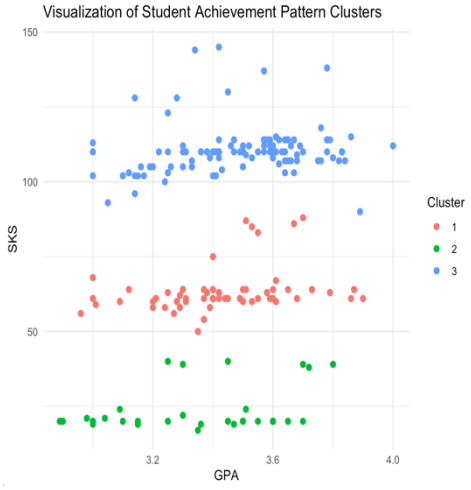


Figure 2. Scatter plot showing clusters of students based on GPA and SKS.

As shown in Figure 2, each point represents a student, colored by their respective cluster. The horizontal spread reflects GPA, while the vertical axis shows total SKS taken. The clusters exhibit the following patterns:

a. Cluster 1 (Red): This group occupies the middle range in both GPA and credit load, representing students with moderate academic standing. Their progression is relatively stable but not exceptional.
b. Cluster 2 (Green): Students in this cluster exhibit both low GPA and low SKS. These may include early-year students, those on academic leave, or those facing performance challenges.
c. Cluster 3 (Blue): Students in this cluster have both high GPA and high credit accumulation. This group likely represents high-achieving students who are not only academically competent but also progressing efficiently through their coursework.

This distribution provides evidence that the clustering structure captures real academic heterogeneity, segmenting students into meaningful categories aligned with their academic progress.

C. Behavioral Profiles of Student Clusters

To complement the academic segmentation, the study also investigated how each cluster differed in learning behaviors, study habits, and motivation, using self-reported indicators from the questionnaire. The average score of each behavioral item was computed per cluster and visualized in the bar chart below.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

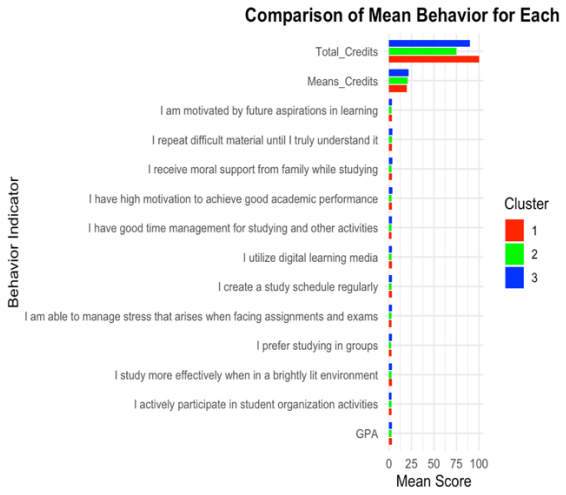


Figure 3. Comparison of behavioral and motivational indicators across student clusters.

From Figure 3**,** it can be observed that:

a. Cluster 1 shows moderate behavioral patterns, consistent with their academic standing. Their scores on indicators such as stress management, digital learning utilization, and goal-setting are average. This suggests a potential to improve outcomes through guided behavioral interventions.
b. Cluster 2 displays lower behavioral and motivational scores. These students report fewer self-regulated learning strategies, lower use of educational technology, and weaker time/stress management. This may explain their low academic performance, highlighting the need for targeted academic coaching and psychological support.
c. Cluster 3 not only excels academically but also scores highest across almost all behavioral indicators. These students demonstrate strong intrinsic motivation (e.g., high scores in "Saya termotivasi oleh cita-cita masa depan"), effective time management, and good coping strategies. This alignment reinforces the role of non-cognitive factors in academic success.

D. Education and Institutional Implications

The three identified clusters offer actionable profiles for educational stakeholders. Recognizing that students differ not only in achievement but also in behavioral attributes allows institutions to tailor interventions more precisely. For example:

1. Cluster 1 students may benefit from structured learning support, such as workshops on advanced study skills or reflective learning strategies.
2. Cluster 2 students require early identification and possibly academic counselling, stress management support, and personalized learning pathways.
3. Cluster 3 students could be engaged as academic ambassadors, peer mentors, or facilitators in collaborative learning environments.

This study demonstrates that the integration of behavioral, cognitive, and academic variables in clustering analysis enhances the interpretability of student profiles. Moreover, it supports a data-driven approach in designing academic advising systems and policies for student success.The clustering process using k = 3 produced three distinct groups of students based on a combination of numerical and categorical attributes. To gain insight into how the clusters differ.

IV. CONCLUSION

This study applied the K-Prototypes clustering algorithm to group undergraduate students based on a combination of academic performance and behavioral characteristics. By incorporating both numerical attributes (such as GPA and credit load) and categorical factors (such as study program and year of entry), the clustering process provided a nuanced segmentation of student profiles. The optimal number of clusters was determined to be three, as indicated by the elbow method. The resulting clusters revealed distinct academic and behavioral patterns: Cluster 3 represented high-performing, well-regulated students; Cluster 1 included moderately performing students with average learning behaviors; and Cluster 2 comprised students with both low academic progress and weaker behavioral indicators.

These findings underscore the importance of integrating behavioral data into academic analytics. They also highlight the potential for educational institutions to develop personalized support strategies that cater to the specific needs of each student group. In particular, early identification of at-risk students through behavioral profiling can inform timely interventions, while high achievers can be empowered as peer mentors to foster a supportive academic environment. Future research may explore longitudinal tracking of these clusters to evaluate their academic trajectories over time, or extend the model with additional psychological, socio-economic, or engagement-related variables to further refine student typologies.

V. RECOMMENDATIONS

Based on the findings, it is recommended that FMIPA Universitas Negeri Medan implement cluster-based academic support strategies to address students' diverse academic and behavioral characteristics. Students in Cluster 1 may benefit from structured learning assistance and study skills development programs, while students in Cluster 2 should receive more intensive academic guidance, motivational support, and time management training. Meanwhile, students in Cluster 3 could be encouraged to participate in peer mentoring programs and academic enrichment activities.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Furthermore, the integration of K-Prototypes clustering into academic monitoring systems could support more personalized and data-driven student assistance. Future research is recommended to incorporate additional variables, such as socioeconomic background and psychological factors, and to conduct longitudinal analyses to examine changes in student performance over time. Comparative evaluations with other clustering algorithms are also suggested to assess the robustness and effectiveness of the proposed approach.

ACKNOWLEDGMENTS

The authors would like to thank the Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Negeri Medan for its generous financial support in funding this research. The funding provided through LPPM's research grant was essential in carrying out the entire study. We sincerely appreciate LPPM's commitment to advancing scientific research and supporting initiatives that address regional public health challenges.

FUNDING INFORMATION

The authors would like to express their sincere gratitude to Universitas Negeri Medan for providing financial support for this research through the Basic Research Grant Program administered by the Institute for Research and Community Service (LPPM), Universitas Negeri Medan, under Research Contract No. 0085/UN33.8/PPKM/PD/2025. This financial support has contributed significantly to the successful completion of this research.

REFERENCES

S. Aljawarneh, "Using data mining techniques to predict students' academic performance," Education and Information Technologies, vol. 26, no. 3, pp. 2811–2826, 2021.

C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, e1355, 2020.

M. K. Affan et al., "Motivation and learning strategies as predictors of academic performance among university students," International Journal of Instruction, vol. 14, no. 2, pp. 377–392, 2021.

W. Putri and N. P. Hidayat, "Academic performance and its influencing factors: A case study in Indonesian university," Jurnal Pendidikan dan Pengajaran, vol. 54, no. 2, pp. 123–132, 2022.

F. Ariani and R. E. Santoso, "A multivariate study on academic burnout, motivation, and performance," International Journal of Learning and Teaching, vol. 16, no. 1, pp. 58–65, 2023.

T. H. Nguyen, T. T. Nguyen, and H. D. Pham, "Student academic performance prediction using machine learning algorithms," Procedia Computer Science, vol. 198, pp. 436–442, 2022.

M. C. Roy and A. Chatterjee, "Machine learning techniques in student performance prediction: A review," Education and Information Technologies, vol. 28, pp. 2457–2478, 2023.

L. G. Trujillo-Torres et al., "Application of clustering techniques in educational environments: A systematic review," Computers in Human Behavior Reports, vol. 5, 100118, 2022.

Z. Khan and R. A. Khan, "A comparative study of clustering algorithms for educational data analysis," Procedia Computer Science, vol. 190, pp. 504–509, 2021.

Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283–304, 2019.

K. W. Lim, M. R. Hassan, and N. R. Ismail, "Predictive analytics for student performance based on behavioral and academic data," Journal of Theoretical and Applied Information Technology, vol. 98, no. 14, pp. 2785–2796, 2020.

J. Almeda, T. Capistrano, G. Sarte, 2010, Elementary Statistics, Quezon City: UP Press.

H. Prasetyo, "Regional clustering based on economic potential with a modified fuzzy k-prototypes algorithm for village developing target determination", Jurnal Teknologi dan Sistem Komputer, vol.10 no.1, 2022, 46-52

J. Quach, R. Malekian "Exploring the Weather Impact on Bike Sharing Usage Through a Clustering Analysis", Journal of Computers Vol. 33 No. 5, 2022, pp. 163-173

S. Chen et al., "A Novel Machine Learning Framework for Comparison of Viral COVID-19–Related Sina Weibo and Twitter Posts: Workflow Development and Content Analysis", JMIR Publications Vol 23 No 1, 2021.

J. Ji et al., "An improved k-prototypes clustering algorithm for mixed numeric and categorical data", Neurocomputing, Vol 120, 2013, 590-596.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.