Clustering Students Based on Academic Performance and Social Factors: An Unsupervised Learning Approach to Identify Student Patterns

Felinda Aprilia Rahma[1,*], Siti Zayyana Ulfah[2]

[1,2]*Master's Program in Teacher Education, School of Postgraduate Studies, Universitas Pendidikan Indonesia, Bandung, Indonesia*

(Received: March 1, 2025; Revised: April 5, 2025; Accepted: July 25, 2025; Available online: September 10, 2025)

Abstract

This study explores the application of K-Means clustering, an unsupervised learning method, to group students based on academic performance and social factors. The primary objective is to uncover hidden patterns among students by analyzing academic scores in mathematics, reading, and writing, as well as demographic attributes including gender, ethnicity, parental education level, and lunch type. Data preprocessing steps, such as normalization and one-hot encoding, were conducted to prepare the dataset for clustering. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, with K=3 selected for its balance between cluster quality and interpretability. The clustering results revealed three distinct groups of students: low performers, average performers, and high performers. These clusters were visualized using PCA and t-SNE, which showed clear separation and internal consistency. Interpretation of the clusters suggests that social factors may influence academic outcomes, with students from disadvantaged backgrounds more likely to fall into the lower-performing group. The study highlights the importance of data-driven approaches in understanding student diversity and designing targeted interventions. Furthermore, this research underlines the potential of clustering techniques to inform educational strategies by identifying students' needs more precisely. However, limitations include reliance on academic and basic demographic variables, and sensitivity of the K-Means algorithm to outliers and the predefined number of clusters. Future research should incorporate additional factors such as emotional well-being and learning preferences to develop more comprehensive educational models. Overall, the study demonstrates that clustering can serve as a valuable tool for enhancing the effectiveness and equity of educational programs.

Keywords: Academic Performance, Educational Interventions, K-Means clustering, Social Factors, Student Grouping

1. Introduction

The organization of students into groups based on academic performance and social factors plays a pivotal role in enhancing educational outcomes. Research has shown that collaborative learning environments foster higher academic engagement, as students who receive strong social support tend to be more motivated and actively participate in their studies. This heightened engagement ultimately leads to improved academic performance [1]. Such social support is particularly valuable during transitional periods, such as the shift from high school to college, where students often experience anxiety. It not only boosts motivation but also facilitates the development of self-efficacy, further contributing to academic success [1], [2].

Furthermore, variations in academic achievement often reflect underlying sociocultural influences and classroom dynamics. Students who form strong connections with their peers and engage meaningfully in cooperative settings are more likely to thrive academically [3]. The classroom environment plays a crucial role in shaping these interactions, which in turn bolster the motivational factors necessary for academic success [4]. By strategically grouping students based on these academic and social characteristics, educators can cultivate supportive peer relationships, implement tailored instructional methods, and create a more enriching learning atmosphere. This approach, in turn, promotes holistic academic success [5].

The growing diversity in student abilities and backgrounds necessitates an approach to education that recognizes and accommodates this variation. In today's educational systems, significant disparities in academic performance are driven

[*] **Corresponding author: Felinda Aprilia Rahma (felindaaprilia16@upi.edu)**

DOI: https://doi.org/10.47738/ijaim.v5i3.109

by a complex mix of sociocultural factors, cognitive styles, and emotional needs. Addressing these differences requires educational strategies that are flexible enough to engage and support students from various demographics. The integration of customized pedagogical approaches that acknowledge each student’s individual strengths and challenges helps create a more inclusive educational environment, ensuring that all students have an equal opportunity for success.

Moreover, fostering effective interpersonal relationships within educational settings is key to bridging the gaps in learning outcomes. Strong teacher-student and peer connections are fundamental for enhancing academic motivation and improving performance [6]. In addition, incorporating diverse teaching resources, such as open educational resources (OER), can reduce barriers created by socio-economic factors, further promoting a more equitable learning environment [7]. By addressing the varied needs of students, educational institutions can build supportive and effective learning climates that serve all students, regardless of their background [6], [7].

In light of these challenges, clustering techniques, especially those based on unsupervised learning, offer powerful methods for uncovering hidden patterns within student data. By grouping students according to their academic performance, behaviors, and social factors, researchers can gain a deeper understanding of the dynamics influencing academic success. For instance, Khamis et al. used clustering to categorize students based on their internet usage and academic records, revealing distinct student profiles that correlated with performance [8]. This approach demonstrates the effectiveness of clustering in identifying patterns that may not be immediately visible through traditional analysis.

Similarly, clustering can also be used to improve educational recommendations, such as personalized learning resources and teaching strategies [9]. The application of K-Means clustering in educational contexts has proven effective in categorizing students based on academic performance, shedding light on disparities in educational outcomes and highlighting potential interventions [10]. Furthermore, clustering techniques have been successfully employed to extract valuable insights from learning management system data, reinforcing their versatility in educational applications [11]. This not only helps in organizing student data but also plays a crucial role in refining educational strategies through data-driven insights.

The contemporary educational landscape is characterized by considerable variations in student academic performance, which educational systems often struggle to address effectively. These disparities arise from multiple factors, including socio-economic background, learning styles, and institutional support systems [12]. Marginalized groups, in particular, face additional barriers such as feelings of inferiority and tokenism, which only serve to deepen performance gaps [12]. Responding to the needs of diverse student populations requires tailored approaches that address these underlying issues. Leveraging techniques like clustering offers a promising avenue to uncover hidden patterns within student data, allowing for a more nuanced classification of students based on their unique characteristics [10].

By employing these methods, educators can more effectively customize learning strategies and resources to meet the varied needs of students, ultimately improving academic outcomes for all learners [13]. Data-driven approaches can help create a more equitable education system and foster improved performance across different student groups [14]. However, the lack of comprehensive and well-analyzed data often complicates efforts to effectively group students based on relevant social and academic factors. Insufficient data collection and analysis hinder the formation of nuanced understandings of student groups, thus undermining efforts to provide targeted interventions [15].

For example, while data-driven models can analyze student interactions and behaviors within educational contexts, their success depends on the quality and completeness of the data inputs. If critical variables are overlooked, it becomes difficult to identify patterns that can guide interventions aimed at addressing students' diverse needs [16]. By incorporating a variety of data sources, such as student feedback and demographic information, the clustering process can be enriched, providing a more accurate and comprehensive view of student groups and thereby enhancing the support provided [17]. Therefore, improving data utilization is essential to developing inclusive educational environments that meet the needs of all students.

Designing educational programs that effectively address the diverse needs of students is an ongoing challenge in modern education. This complexity arises from the diverse backgrounds and learning requirements of students, necessitating a deep understanding of their unique characteristics [18]. For instance, research by Arifin et al. highlights the gap between existing curricula and the needs of students in visual communication design, underscoring the importance of aligning educational content with students' expectations and experiences [18]. Achieving such alignment requires continuous analysis and adaptation of instructional materials and pedagogical approaches.

Tailored educational programs depend on accurately identifying students' learning preferences and behavioral patterns. Research shows that clustering can group students based on shared attributes, providing valuable insights into their academic needs. This data-driven approach not only improves the effectiveness of educational interventions but also empowers educators to implement strategies that enhance engagement and success across diverse student demographics [19]. By integrating comprehensive data analytics into educational program design, institutions can more effectively address the specific needs of diverse student groups, leading to a more equitable and effective educational landscape [20].

2. Literature Review

2.1. Importance of Student Grouping

The importance of strategic student grouping based on academic and social performance has been well-documented in enhancing educational interventions. Group work, particularly in collaborative learning environments, has shown to significantly improve academic outcomes by fostering peer interaction, idea-sharing, and critical thinking. Studies like those by Saputro [21], emphasize the value of group work in fields such as nursing education, where peer discussions and debates help develop clinical skills and deepen students' understanding of complex concepts. Cooperative learning models enhance both academic skills and students' self-esteem, fostering social and academic growth. Through such interactions, students are encouraged to view challenges from multiple perspectives, contributing to a more holistic educational experience.

The effectiveness of collaborative teaching models like the Group Investigation (GI) approach. Their research reveals that learning methods which promote student interaction, such as GI, help students develop a deeper understanding of complex subjects. Additionally, engaging instructional methods boost students' motivation and conceptual grasp, leading to improved performance. These findings highlight that well-structured, interaction-based frameworks not only enhance academic outcomes but also stimulate intrinsic motivation, fostering a deeper engagement with the material.

However, challenges such as social loafing, where students disengage in group tasks, can hinder the effectiveness of group-based learning. Research by Riwoe et al. [22] suggests that group dynamics must be carefully managed to ensure full participation and accountability. To optimize academic outcomes, Cvitković et al. [23] argue that educators must address the diverse needs and capabilities of students within groups, fostering an environment that encourages active participation and boosts academic self-efficacy. In conclusion, strategic student grouping plays a pivotal role in fostering effective learning environments, with continuous adaptations in teaching methods necessary to overcome challenges and maximize the benefits of collaborative learning.

2.2. Clustering Techniques in Education

Clustering techniques, particularly K-Means, have gained widespread application in educational research for analyzing student performance and social factors. These algorithms help categorize students based on various attributes, enabling more personalized educational interventions and improving overall learning outcomes. K-Means clustering to optimize student grouping based on learning styles in programming classes, cluster analysis can reveal patterns in student engagement and learning behaviors, allowing educators to adjust their teaching strategies to better accommodate diverse student populations.

In addition to K-Means, other clustering methods, such as hierarchical clustering, are also valuable in educational settings. Lorente-Echeverría et al. [24] used this approach to analyze teachers' profiles in relation to sustainability practices in physical education. By evaluating cluster cohesion and separating subgroups based on squared Euclidean distances, they demonstrated how such a structured clustering analysis could be applied to student data as well. This methodology ensures clear boundaries between groups, which can be instrumental in precisely targeting educational strategies and resources. Similarly, Ren [25] discusses how clustering techniques optimize educational resources by aligning subject matter relevance and learning objectives, further enhancing the quality of educational delivery.

Moreover, the versatility of clustering extends to diverse educational contexts. Ligibel et al. [26] illustrate how cluster randomization can be used to analyze students' responses to physical activity interventions, highlighting its effectiveness in clinical educational settings. The integration of clustering with decision trees in educational data mining, focusing on predicting student performance based on self-reported learning patterns. This approach reinforces the utility of clustering in identifying students who require additional support. In conclusion, clustering techniques, such as K-Means, provide invaluable insights into student performance and social dynamics, fostering data-driven educational practices that can improve academic outcomes and overall student well-being.

2.3. Social Factors Influencing Student Performance

Social factors such as gender, ethnicity, parental education, and socioeconomic status play a significant role in shaping student performance. Research underscores the importance of these factors in influencing academic outcomes and highlights the need for educators, policymakers, and researchers to address these dynamics in order to foster more equitable learning environments. For example, gender and ethnicity have been found to be critical in determining academic success. Hwang et al. [27] found that matching teacher and student ethnicity, especially in elementary settings, can lead to better academic performance. Similarly, Glock et al. [28] suggest that ethnic minority students may perceive less support from teachers, which can negatively impact their achievement. This points to the need for more diverse teaching staff to create an inclusive environment that supports all students.

Parental education is another influential factor, as studies show that students' perceptions of their parents' emotional support and the level of parental education correlate with better academic performance. Liu et al. [6] highlighted that students with higher parental educational attainment tend to have more academic support, leading to better outcomes. Gul et al. [29] further reinforce this by emphasizing the role of parental involvement in academic achievement. Additionally, socioeconomic factors, such as lunch type, have a profound impact on student performance. Segkulu [30] notes that students from lower socioeconomic backgrounds often struggle due to limited access to resources, which impedes their academic engagement. The type of lunch students receive can affect their energy and focus, with those receiving adequate nutrition showing better academic capabilities.

Finally, social engagement also plays a vital role in academic success. Abubakar et al. [31] found that students who participate actively in school activities tend to perform better academically, regardless of their social background. This highlights the importance of creating a positive school climate and encouraging peer interactions to support academic achievement. In conclusion, understanding the influence of social factors such as gender, ethnicity, parental education, and nutrition is crucial for educators and stakeholders to implement targeted strategies that promote equitable learning opportunities and improve student outcomes.

2.4. Strengths and Limitations of K-Means Clustering

K-Means clustering has become a widely used technique in educational research due to its simplicity, computational efficiency, and ability to organize large datasets into meaningful clusters. It allows educators and researchers to analyze student performance, engagement, and other educational metrics effectively. One of the primary strengths of K-Means, as highlighted by Faisal et al. [32], is its ability to partition data based on similarities, enabling the identification of patterns in student competencies. The algorithm minimizes the Sum of Squared Errors (SSE) iteratively, which makes it adaptable across various educational contexts. For example, Agha et al. [20] demonstrated the utility of K-Means in forecasting academic performance by uncovering underlying trends in educational data. This ability to reveal patterns can help institutions make data-driven decisions that ultimately benefit student outcomes. Additionally, the algorithm's flexibility allows it to be used in diverse scenarios, from grouping students based on academic performance to identifying trends in engagement levels [33], making it valuable for addressing individual learning needs.

However, K-Means clustering does have notable limitations. One significant drawback is the requirement to predefine the number of clusters (k), which can greatly affect the results of the analysis. The choice of k is often arbitrary, and an incorrect selection can lead to poor cluster formation and misinterpretation of the data [6], [33]. Liu et al. [6] caution that incorrect assumptions about the number of clusters can skew analysis, particularly when assessing student data. Another limitation is K-Means' sensitivity to outliers and noise within the data. Additionally, the reliance on Euclidean distance for measuring similarity may not always be appropriate, especially for data with varying scales and distributions, as noted by Agha et al. [20].

Furthermore, K-Means can yield inconsistent results when working with high-dimensional datasets [34]. This issue may require dimensionality reduction techniques to improve clustering performance, adding complexity to the analysis process. In comparison to other clustering methods like K-Medoids or Hierarchical clustering, K-Means may lack robustness, particularly when clusters are not spherical or evenly sized [35]. These limitations can affect the accuracy of the results and the strategies implemented based on the clustering outcomes. In conclusion, while K-Means clustering offers significant strengths in computational efficiency and flexibility for educational data analysis, it also has limitations that must be carefully managed. Understanding both the strengths and weaknesses of K-Means is essential for researchers and educators looking to use clustering techniques effectively to support student learning and improve academic performance.

3. Methodology

The following diagram illustrates the steps involved in the data analysis process using the K-Means algorithm, as shown in figure 1.

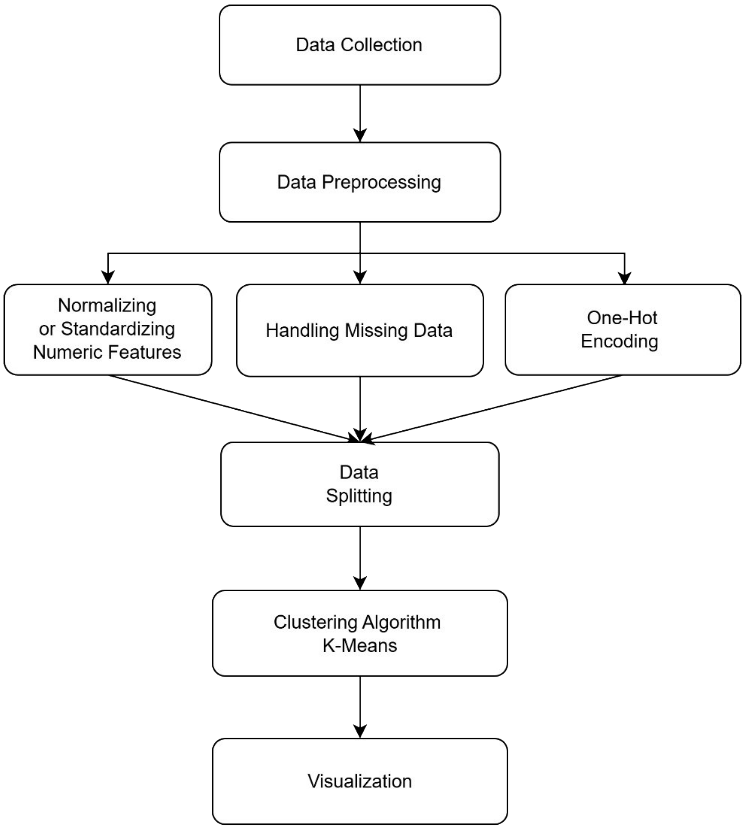

Figure 1. Research Methodology

3.1. Data Collection

The research object in this study consists of students whose data includes academic performance in mathematics, reading, and writing, as well as key social features. These social factors will be used as features for clustering the students. Table 1 summarizing the features considered for this research:

Table 1. Student Academic and Social Features

Feature	Description
Math Score	Student's score in mathematics.
Reading Score	Student's score in reading.
Writing Score	Student's score in writing.

Gender	The gender of the student (Male/Female).
Ethnicity	The ethnic group of the student (e.g., Group A, Group B, etc.).
Parental Education Level	The highest level of education attained by the student's parents (e.g., High School, Bachelor's Degree, etc.).
Lunch Type	Whether the student receives a standard or free/reduced lunch.

These features will serve as the basis for clustering the students. By considering both academic and social factors, this research aims to uncover clusters of students who share similar patterns in their academic performance and social backgrounds. The findings from these clusters could provide valuable insights into the influence of social factors on academic performance and help inform the design of targeted educational interventions for different student groups.

3.2. Data Preprocessing

The data preprocessing steps are crucial for ensuring the quality and suitability of the data for clustering analysis. The first step involves data processing, which includes identifying and handling any missing values. Missing data can be addressed by either imputing values (using mean, median, or mode) or removing records with significant missing data, depending on the context. Next, one-hot encoding is applied to the categorical variables, such as gender, ethnicity, parental education level, and lunch type, to convert them into numeric representations. This transformation is necessary because clustering algorithms like K-Means can only work with numerical data. For instance, gender might be encoded as 0 for male and 1 for female, and different ethnic groups might be encoded as separate binary variables.

Once the categorical variables are encoded, the numeric features such as math scores, reading scores, and writing scores are normalized or standardized. Normalization or standardization ensures that the scales of the different numeric variables are similar, preventing any one feature from disproportionately influencing the clustering process. For example, without scaling, math scores with a much larger range than writing scores might dominate the clustering process. The StandardScaler method is commonly used to standardize features so that they have a mean of 0 and a standard deviation of 1.

Finally, data splitting is performed to organize the data for clustering. The dataset is divided into relevant features for clustering, ensuring that the social and academic data are in formats that can be analyzed. The clustering process will consider all these features together, allowing for a holistic understanding of how academic and social characteristics interact in determining the student groups. Once the preprocessing is complete, the data will be ready for input into the K-Means Clustering algorithm, where the features will be grouped into clusters based on their similarities.

3.3. Clustering Algorithm

K-Means Clustering is employed as the primary algorithm to group students based on their academic and social characteristics. The K-Means algorithm is a popular unsupervised learning method that partitions the data into K clusters, where each cluster consists of students with similar attributes. The algorithm minimizes the sum of squared distances between data points and their respective cluster centroids. Specifically, for each data point x_i, the algorithm assigns it to the cluster C_k who’s centroid μ_k is the closest, according to the following formula:

$$\text{argmin}_k \sum_{i=1}^{n} \|x_i - \mu_k\|^2 \tag{1}$$

Note:
x_i is the i-th data point.
$\mu \;_k$ is the centroid of the k-th cluster.
n is the total number of data points.

To determine the optimal number of clusters (K), the Elbow Method and Silhouette Score are used.

The Elbow Method helps identify the point at which adding more clusters does not significantly improve the model. The method involves plotting the inertia (sum of squared distances from each point to its assigned centroid) for different values of K. The inertia for a given number of clusters K is given by:

$$\text{Inertia}\,(K) = \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik} \|x_i - \mu_k\|^2 \tag{2}$$

Note:

r_{ik} is the membership indicator (1 if point x_i belongs to cluster k, 0 otherwise).

μ_k is the centroid of cluster k.

The optimal number of clusters is chosen where the elbow in the inertia plot appears, this is where adding more clusters no longer results in a significant decrease in inertia.

The Silhouette Score is used to evaluate the separation of the clusters. It measures how similar each data point is to the other points in its cluster compared to points in the nearest neighboring cluster. The Silhouette Score for a given data point x_i is computed as:

$$s\left(x_i\right) = \frac{b\left(x_i\right) - a(x_i)}{\max\left(a\left(x_i\right), b(x_i)\right)} \tag{3}$$

Note:

$a\left(x_i\right)$ is the average distance between x_i and all other points in the same cluster (intra-cluster distance).

$b\left(x_i\right)$ is the minimum average distance between x_i and all points in any other cluster (inter-cluster distance).

The overall Silhouette Score is the average of the individual Silhouette Scores for all points in the dataset, and a higher value indicates better-defined clusters. The Silhouette Score ranges from -1 to 1, with a higher value indicating well-separated and appropriately assigned clusters.

Once the optimal number of clusters is determined using the Elbow Method and Silhouette Score, the K-Means algorithm is applied to the dataset. After clustering, model validation is performed using the Silhouette Score to assess the quality of the clustering results. A higher Silhouette Score indicates that the clusters are well-separated and that each student is appropriately grouped based on their academic and social features. This validation step ensures that the clustering results are meaningful and provide useful insights into the data.

3.4. Visualization

After performing the clustering, it is important to visualize the resulting clusters to better understand the distribution and relationships between the data points. In this study, scatter plots and dimensionality reduction techniques, such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), are used to display the cluster distribution in a 2D or 3D space.

Scatter plots are a simple and effective way to visualize clusters, especially when dealing with two or three features. In this case, we can plot the data points from the clusters using math score vs. reading score or writing score. Each data point in the scatter plot will be color-coded according to its assigned cluster label, allowing for an easy visual identification of how the students are grouped based on their academic performance.

PCA is a linear dimensionality reduction technique that projects the data onto the top two or three principal components, capturing the most variance in the data. PCA is useful for visualizing high-dimensional data by reducing it to lower dimensions without losing much information. After performing PCA, the data can be plotted in 2D or 3D space, where each data point is colored according to its cluster label. The resulting plot allows for a visual understanding of how well the clusters are separated along the principal components. The formula for PCA is as follows:

$$Z = XW \tag{4}$$

Z is the matrix of principal components (lower-dimensional representation), X is the original data matrix, and W is the matrix of eigenvectors (the directions of maximum variance).

t-SNE is a non-linear dimensionality reduction technique that is particularly well-suited for visualizing high-dimensional datasets in 2D or 3D space. t-SNE minimizes the divergence between probability distributions in high-dimensional space and their 2D or 3D counterparts, allowing it to group similar points together. This technique is effective for visualizing complex clusters when the data has more than two dimensions. It is particularly useful when the clusters might not be linearly separable. The formula for t-SNE involves pairwise affinities between data points and their representation in lower dimensions, but for simplicity, it's mostly used via libraries like scikit-learn.

The visualizations using PCA and t-SNE help in identifying whether the clusters are well-separated, whether there is any overlap between groups, and whether there are any outliers or unusual patterns in the data. These visual representations provide an intuitive way to interpret the clustering results and assess the effectiveness of the grouping algorithm.

4. Results and Discussion

4.1. Result

Figure 2 shows the silhouette scores for various values of K tested in this analysis. The silhouette score is used to evaluate how well each object is grouped within its assigned cluster, with higher scores indicating better separation between clusters. At K=7, the silhouette score reaches its highest value of approximately 0.149, suggesting that the clustering at this point exhibits the best separation.

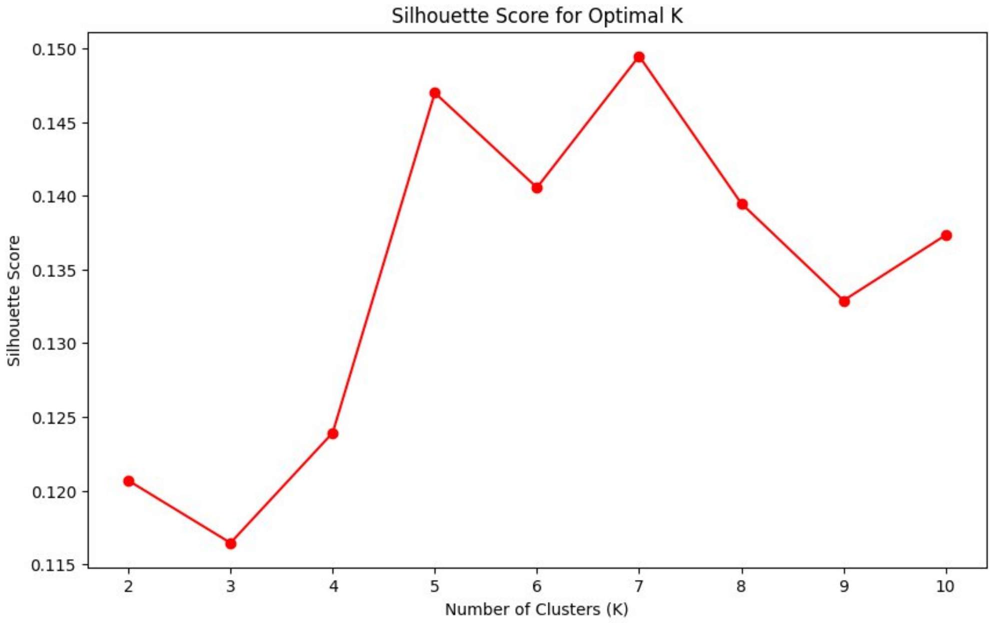

Figure 2. Silhouette Scores

However, further analysis reveals that K=3 and K=5 also yield consistent and interpretable results, with clear separation between clusters. K=3 was ultimately selected as the optimal number of clusters because, in addition to providing a reasonable level of separation, it offers simpler and more practically relevant interpretation. A smaller number of clusters is easier to analyze and explain. While several values of K provide good results, K=3 was considered the best choice for this analysis due to the balance it strikes between cluster separation quality and ease of interpretation.

The optimal number of clusters for the dataset was determined using the Elbow Method, as shown in figure 3. In this method, the inertia (sum of squared distances between points and their cluster centers) is plotted against the number of clusters. The plot reveals a clear elbow at K=3, where the rate of decrease in inertia slows down significantly. Before K=3, adding more clusters resulted in a substantial reduction in inertia, indicating that the clusters were becoming more compact. However, beyond K=3, the reduction in inertia becomes less pronounced, and the improvements in clustering quality become marginal. This flattening of the curve suggests that additional clusters do not add substantial value in terms of better grouping, making K=3 the most appropriate and meaningful choice for the dataset.

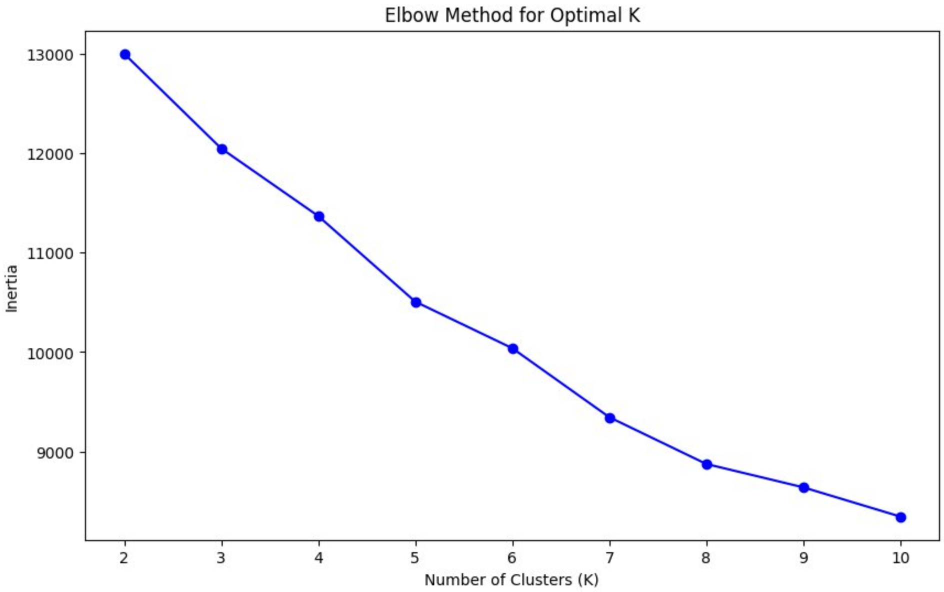

Figure 3. Elbow Method

Following the identification of the optimal number of clusters, the three clusters revealed distinct patterns in academic performance. It is important to interpret these clusters not only based on academic scores but also considering potential underlying social factors. Cluster 0 (Purple), which contains students with lower scores in both math and reading, may represent a group that faces various challenges. These students might come from socioeconomically disadvantaged backgrounds, where access to educational resources or family support is limited. Additionally, students in this cluster might struggle with motivation or face environmental stressors, such as language barriers or insufficient academic guidance at home. This cluster may also be overrepresented by students from specific ethnic or cultural groups who historically face educational inequities.

Cluster 1 (Teal), which includes students with average academic performance, may represent a more balanced group in terms of academic abilities. Students in this cluster might come from families with moderate socioeconomic backgrounds, where they receive a reasonable amount of educational support but lack access to more specialized resources or enrichment programs. It’s possible that this group includes a mix of students from diverse ethnic backgrounds, with a variety of academic experiences that result in a middle-range performance. These students might be in need of some academic enrichment or intervention to help boost their performance and push them to higher levels.

Cluster 2 (Yellow), consists of students who show higher scores in both math and reading, identifying them as high performers. This group may include students from more privileged backgrounds, where access to high-quality educational resources, parental support, and extracurricular activities is more common. These students could also come from ethnic groups that are more likely to have educational support systems in place, which foster academic success. Additionally, students in this cluster may benefit from more advanced learning opportunities, and their higher performance may be attributed to the availability of enrichment programs or a more robust academic environment both at home and in school.

To provide a deeper understanding of how social factors, such as ethnicity, family income, and parental education levels, influence academic performance, further analysis would be valuable. For instance, by examining how students from specific ethnic backgrounds or low-income families are distributed across the clusters, we could identify any disparities and understand how these factors might contribute to the observed differences in academic performance.

To enhance this analysis, t-SNE (t-Distributed Stochastic Neighbor Embedding) is used as a dimensionality reduction technique to visualize high-dimensional data in a lower-dimensional space, shown in figure 4. The three clusters are clearly separated, indicating that there is a significant differentiation between the groups based on their academic performance. The well-defined boundaries between the clusters suggest that the clustering algorithm has successfully identified distinct patterns within the data.

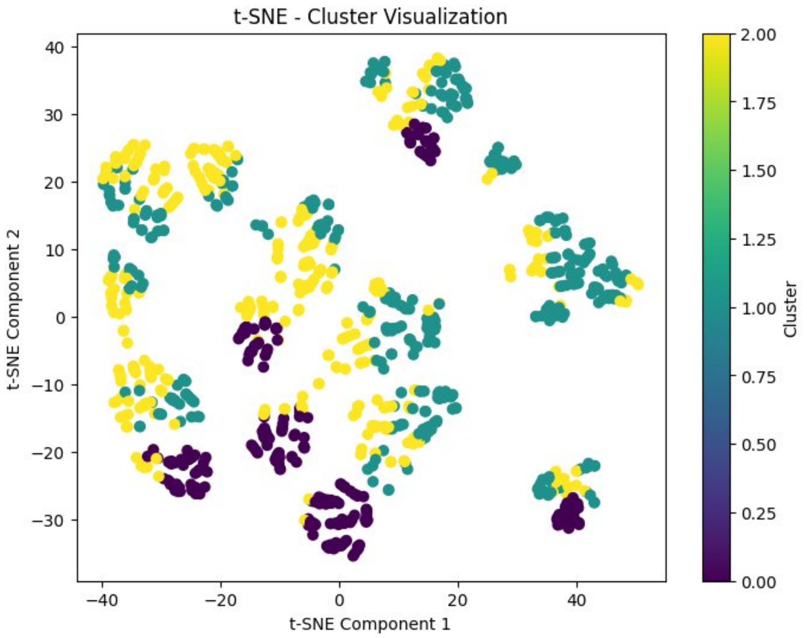

Figure 4. t-SNE Cluster Visualization

Moreover, the tightness of each cluster in the t-SNE plot indicates that students within each group share similar academic performance characteristics, supporting the idea that students in the same cluster are academically similar to one another.

In addition to t-SNE, PCA further validates the clustering process, as shown in figure 5. The clusters are primarily separated along the first principal component, suggesting that the largest variance in student performance is captured by this component. This confirms that the features contributing to student performance, such as math, reading, and writing scores, are strongly aligned with the principal components that explain the major variations in performance across the groups.

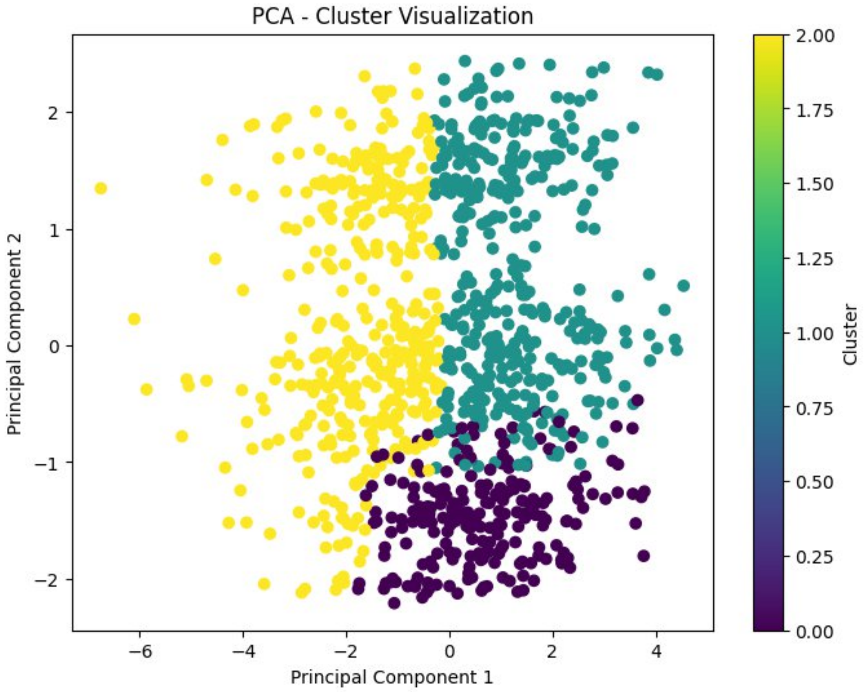

Figure 5. PCA Cluster Visualization

The fact that the clusters are well-separated in the PCA plot indicates that PCA has effectively reduced the complexity of the data while retaining the most important information.

Finally, the scatter plot in figure 6 compares math and reading scores, showing how students are grouped into three clusters based on their academic performance. Cluster 2 (Yellow) is in the top-right corner, indicating high scores in both subjects, representing high performers. Cluster 1 (Teal) is in the middle, showing average performance, while Cluster 0 (Purple) is in the lower-left, indicating lower scores. The centroids, or average positions, of each cluster further define the groups. Cluster 0 (Purple) has lower average scores, suggesting these students would benefit from academic support. Cluster 1 (Teal), with average performance, could benefit from enrichment programs, while Cluster 2 (Yellow), with higher scores, may be suited for advanced academic opportunities.

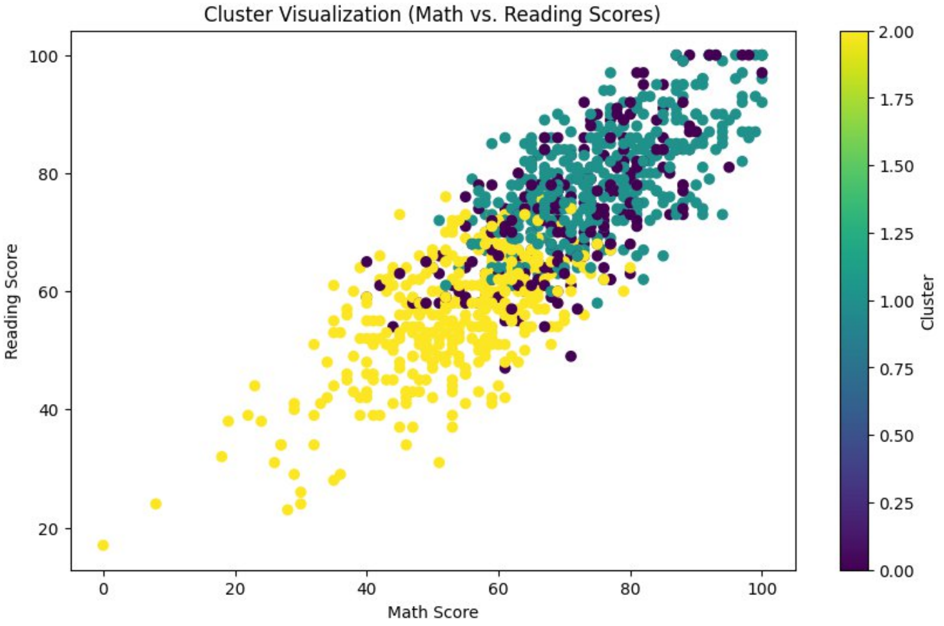

Figure 6. Cluster Visualization (Math vs. Reading Scores)

The clustering analysis has successfully segmented students into three distinct groups based on their academic performance. Moving forward, targeted interventions, enrichment opportunities, and advanced programs can be implemented to address the needs of students at different academic levels. Students in Cluster 0 (Purple), who exhibit lower performance in both math and reading, would benefit from additional academic support or remediation strategies aimed at addressing their specific learning gaps. Cluster 1 (Teal), with middle-range performance, could thrive with enrichment or supplemental academic programs designed to challenge them and encourage further academic growth. Meanwhile, students in Cluster 2 (Yellow), who demonstrate above-average performance, should be offered advanced academic opportunities, such as gifted programs or specialized learning tracks, to maximize their potential. Further analysis incorporating demographic and social factors, such as ethnicity, socioeconomic status, and parental education, would offer a more comprehensive understanding of the factors influencing student performance and help refine the interventions to ensure they are tailored to the unique needs of each group.

4.2. Discussion

The results of the clustering analysis provide valuable insights into how students can be grouped based on their academic performance, using various clustering techniques to enhance understanding. Figure 2, which shows the silhouette scores for different values of K, reveals that while K=7 offers the highest silhouette score of approximately 0.149, further examination highlights that K=3 and K=5 also yield consistent and meaningful results. Although K=7 provides the best silhouette score, K=3 was ultimately chosen as the optimal number of clusters due to its balance between cluster separation quality and interpretability. This smaller number of clusters is easier to analyze and provides a more practical and relevant interpretation for educators seeking to segment students based on their academic performance. The silhouette score, which evaluates the separation between clusters, indicates that the clustering at K=3 provides a sufficiently clear distinction between groups while ensuring simplicity in analysis. This finding aligns with research by Saputro [21], who emphasized that fewer clusters often lead to clearer, more interpretable groupings, making the data more actionable.

In addition to the silhouette analysis, the Elbow Method was used to determine the optimal number of clusters. The plot reveals a clear elbow at K=3, where the rate of decrease in inertia slows significantly. This suggests that adding more clusters beyond K=3 does not result in substantial improvements in clustering quality. Before K=3, the addition of clusters reduces inertia considerably, signifying stronger grouping. However, beyond K=3, the improvements become marginal, confirming that K=3 is the most meaningful and appropriate choice. These methods together reinforce the selection of K=3 as the optimal number of clusters for this dataset, similar to the findings of Ren [25], who also identified that K=3 yielded the most meaningful separation in student performance datasets.

The three identified clusters demonstrate distinct patterns in student academic performance. Cluster 0 (Purple), consisting of students with lower scores in both math and reading, likely represents a group facing challenges such as limited resources, lack of academic support, or external stressors such as language barriers. This group may be overrepresented by students from socioeconomically disadvantaged backgrounds, where access to educational resources is restricted. Research by Gul et al. [29] found that students from lower socioeconomic backgrounds often face barriers that hinder their academic success, which supports the interpretation of Cluster 0 as a group needing additional academic support. Cluster 1 (Teal), with average performance in both subjects, may represent students from more balanced backgrounds who lack advanced resources but still receive adequate support. These students could benefit from enrichment programs or additional academic interventions. Cluster 2 (Yellow), representing high performers, includes students with above-average scores in both subjects. This group may come from more privileged backgrounds with better access to educational resources and support systems. Research by Liu et al. [6] reinforces this interpretation, suggesting that students from higher socioeconomic backgrounds tend to perform better academically due to greater parental support and access to resources.

This analysis sheds light on how academic performance can be used to effectively group students into meaningful clusters, helping to identify distinct groups that require different levels of educational intervention. By employing a range of clustering techniques such as K-means, t-SNE, and PCA, the results offer a clearer picture of the varying academic needs of students, making it possible to implement more targeted, data-driven educational strategies. This approach can also help optimize the use of resources by focusing on the specific needs of each group, ultimately supporting better educational outcomes.

Understanding the academic grouping of students holds significant potential for shaping educational strategies that are more responsive to student needs. Students in Cluster 0 (Purple), who show lower academic performance, are in urgent need of focused remediation and tailored support programs to help them overcome learning barriers. Cluster 1 (Teal), with average performance, would benefit from enrichment programs that provide additional challenges and opportunities for growth. For Cluster 2 (Yellow), composed of high-performing students, offering advanced learning opportunities, such as specialized tracks or gifted programs, would help further cultivate their potential.

Moreover, the findings highlight the critical need to consider social factors, such as socioeconomic status and access to resources, when designing interventions. Hwang et al. [27] suggested that addressing these factors can significantly impact student success, and the current results underscore the importance of tailoring educational policies and practices to support students based on their specific academic and social contexts. By taking into account these additional dimensions, educators can create more equitable and effective learning environments, ensuring that all students receive the support they need to succeed.

However, there are several limitations to this analysis. First, the clustering analysis relies solely on academic performance data (math, reading, and writing scores), which may not fully capture all the factors that influence student achievement. Variables such as student motivation, emotional well-being, and external support factors, like family involvement or social environment, are not considered here but could significantly impact performance. For a more comprehensive understanding, future studies could integrate these additional factors into the analysis, as Cvitković et al. [23] suggested that a more holistic view of student performance should include emotional and social factors alongside academic ones.

Another limitation is the use of K-means clustering, which requires the predefined selection of the number of clusters (K). Although K=3 was chosen as the optimal number, the process of selecting K is inherently subjective and can influence the results. Furthermore, K-means is sensitive to outliers, which can distort the clustering outcome. Future analyses could explore other clustering methods, such as K-medoids or hierarchical clustering, which may offer greater robustness in handling outliers and different data distributions.

Finally, the dimensionality reduction methods, such as t-SNE and PCA, are effective in visualizing the clusters, but they do have limitations in fully capturing the complexity of high-dimensional data. While they provide a simplified view, these methods may not always retain all the nuances of the data, and additional validation of the results through other analytical methods would enhance the robustness of the findings. This limitation is acknowledged by Lorente-Echeverría et al. [24] who noted that dimensionality reduction techniques can sometimes oversimplify complex datasets, which may affect the overall accuracy of the clustering results.

In conclusion, the clustering analysis successfully segmented students into three distinct groups based on their academic performance, with clear visualizations and centroid analysis revealing meaningful differences between the clusters. The findings suggest targeted interventions for each group, with Cluster 0 (Purple) needing additional academic support, Cluster 1 (Teal) benefiting from enrichment programs, and Cluster 2 (Yellow) requiring advanced academic opportunities. However, the analysis also highlights the importance of addressing social factors and acknowledges the limitations of the clustering method used. Further research incorporating more diverse data and exploring alternative clustering methods would help refine these results and ensure that interventions are tailored to the unique needs of each student group.

5. Conclusion

This study demonstrates the effective use of K-Means Clustering to categorize students into distinct groups based on their academic performance and key social factors. By analyzing scores in mathematics, reading, and writing alongside demographic attributes such as gender, ethnicity, parental education, and lunch type, three meaningful clusters were identified: low, average, and high academic performers. These clusters reveal patterns that are not only statistically significant but also practically relevant for understanding student diversity in educational contexts.

The analysis showed that students in the low-performing cluster may face greater academic and social challenges, possibly linked to limited access to resources or support at home. Those in the average-performing cluster could benefit from enrichment programs that build on their existing strengths, while students in the high-performing cluster are strong candidates for advanced academic opportunities. Visualization techniques like PCA and t-SNE further validated the quality of the clusters, demonstrating clear separations and internal consistency among groups, which supports the usefulness of clustering in educational decision-making.

However, several limitations must be acknowledged. The clustering process focused primarily on academic and basic demographic data, without accounting for psychological or emotional variables that can significantly impact student learning. Additionally, K-Means requires the number of clusters to be defined in advance and is sensitive to outliers, which may affect the accuracy of results. Despite these constraints, the study illustrates the value of data-driven approaches in designing more targeted and equitable educational interventions. Future research should incorporate broader datasets and alternative clustering techniques to enhance the depth and impact of such analyses.

6. Declarations

6.1. Author Contributions

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

6.2. Data Availability Statement

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

6.3. Funding

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

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. 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] X. Zhang and W. Qian, "The Effect of Social Support on Academic Performance Among Adolescents: The Chain Mediating Roles of Self-Efficacy and Learning Engagement," *PLoS One*, vol. 19, no. 12, pp. e0311597, 2024, doi: 10.1371/journal.pone.0311597.

[2] T. Xu, P. Zhu, Q. Ji, W. Wang, M. Qian, and G. Shi, "Psychological Distress and Academic Self-Efficacy of Nursing Undergraduates Under the Normalization of COVID-19: Multiple Mediating Roles of Social Support and Mindfulness," *BMC Med. Educ.*, vol. 23, no. May, pp. 1–10, 2023, doi: 10.1186/s12909-023-04288-z.

[3] J. Zúñiga, M. L. Rodríguez-Paz, A. Herrera, and S. Osorio-Toro, "Influence of Sociodemographic Factors on Academic Performance in the Subject of Human Gross Anatomy," *Int. J. Morphol.*, vol. 41, no. 1, pp. 96–103, 2023, doi: 10.4067/s0717-95022023000100096.

[4] B. F. Nwokedi, "Influence of Classroom Environment on the Academic Performance of Students in English Language," *Int. J. Adv. Soc. Sci. Educ.*, vol. 1, no. 4, pp. 191–198, 2023, doi: 10.59890/ijasse.v1i4.732.

[5] T. Tusyanah, E. Handoyo, E. Suryanto, F. R. Indïra, and T. M. Mayasari, "What Affects Students' Academic Performance and Soft Skills Based on the Community of Inquiry (CoI) Theory?," *Int. J. Technol. Educ.*, vol. 6, no. 1, pp. 49–68, 2023, doi: 10.46328/ijte.345.

[6] J. Liu, L. Yu, X. Zhao, Y. Liu, and L. Jiao, "Creativity Profiles and the Role of Interpersonal Relationships in Primary School Pupils: A Person-Centered Approach," *J. Creat. Behav.*, vol. 57, no. 1, pp. 37–48, 2022, doi: 10.1002/jocb.560.

[7] K. E. Ernstmeyer and E. Christman, "Adopting Open Educational Resources as an Equity Strategy," *Nurs. Educ. Perspect.*, vol. 44, no. 5, pp. 306–307, 2023, doi: 10.1097/01.nep.0000000000001170.

[8] S. Khamis, M. Ahmad, A. Ahmad, and M. N. Ahmad, "Internet Use Behaviour Model for Predicting Students' Performance," *Expert Syst.*, vol. 39, no. 8, pp. e12999, 2022, doi: 10.1111/exsy.12999.

[9] A. Andre, N. Suciati, H. Fabroyir, and E. Pardede, "Educational Data Mining Clustering Approach: Case Study of Undergraduate Student Thesis Topic," *IEEE Access*, vol. 11, no. November, pp. 130072–130088, 2023, doi: 10.1109/access.2023.3332818.

[10] A. S. Paramita, "An Unsupervised Learning and EDA Approach for Specialized High School Admissions," *J. Appl. Data Sci.*, vol. 5, no. 2, pp. 316–325, 2024, doi: 10.47738/jads.v5i2.178.

[11] M. C. Carrión, "Data Analysis of Short - Term and Long - Term Online Activities in LMS," *Tem J.*, vol. 11, no. 2, pp. 497–505, 2022, doi: 10.18421/tem112-01.

[12] R. D. Davis and Z. S. Wilson-Kennedy, "Leveraging Cultural Wealth, Identities and Motivation: How Diverse Intersectional Groups of Low-Income Undergraduate STEM Students Persist in Collegiate STEM Environments," *Educ. Sci.*, vol. 13, no. 9, pp. 888, 2023, doi: 10.3390/educsci13090888.

[13] L. K. Smirani, H. A. Yamani, L. Jamel, and J. A. Boulahia, "Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths," *Sci. Program.*, vol. 2022, no. 1, pp. 1–15, 2022, doi: 10.1155/2022/3805235.

[14] R. Kathiah, P. D. A, S. Selvakumar, A. Kunnumbrath, and K. Meenakshisundaram, "Deciphering the Nexus: Exploring Learning Styles and Academic Success Among Medical Students Through a Comprehensive Study," *Cureus*, vol. 16, no. 4, pp. e59079, 2024, doi: 10.7759/cureus.59079.

[15] P. Toukiloglou and S. Xinogalos, "A Systematic Literature Review on Adaptive Supports in Serious Games for Programming," *Information*, vol. 14, no. 5, pp. 277, 2023, doi: 10.3390/info14050277.

[16] I. Mustapha, V. Rattanawiboonsom, and R. Intanon, "Data-Driven Insights in Higher Education: Exploring the Synergy of Big Data Analytics and Mobile Applications," *Int. J. Interact. Mob. Technol.*, vol. 17, no. 20, pp. 21–37, 2023, doi: 10.3991/ijim.v17i20.45037.

[17] M. Li, "Prediction of the Age of Abalones Based on Machine Learning Algorithms," *Appl. Comput. Eng.*, vol. 20, no. 1, pp. 247–255, 2023, doi: 10.54254/2755-2721/20/20231100.

[18] I. Arifin, B. A. Rauf, and A. Arifin, "Analysis of Learning Needs for Visual Communication Design Students for the Latest Graphics Teaching Materials," *Asian J. Educ. Soc. Stud.*, vol. 48, no. 4, pp. 142–153, 2023, doi: 10.9734/ajess/2023/v48i41093.

[19] L. Aguagallo, F. Salazar-Fierro, J. García-Santillán, M. P. Yépez, P. Landeta-López, and I. García-Santillán, "Analysis of Student Performance Applying Data Mining Techniques in a Virtual Learning Environment," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 11, pp. 175–195, 2023, doi: 10.3991/ijet.v18i11.37309.

[20] D. Agha, A. F. Meghji, and S. Bhatti, "Clusters of Success: Unpacking Academic Trends With K-Means Clustering in Education," *VFAST Trans. Softw. Eng.*, vol. 11, no. 4, pp. 15–31, 2023, doi: 10.21015/vtse.v11i4.1633.

[21] S. H. Saputro, "The Impact of Problem Based Learning on Learning Outcomes in Nursing Students," I*nt. J. Multidiscip. Res. Anal.*, vol. 05, no. 10, pp. 2784–2788, 2022, doi: 10.47191/ijmra/v5-i10-29.

[22] C. E. Riwoe, M. D. C. Lerik, and J. M. Y. Benu, "Social Loafing Behavior in Group Task Completion of University Student," *J. Heal. Behav. Sci.*, vol. 4, no. 3, pp. 460–468, 2022, doi: 10.35508/jhbs.v4i3.7330.

[23] D. Cvitković, S. S. Pišonić, and I. Radosevic, "Predictors of Academic Self-Efficacy of University Students: Grades, Learning Disabilities and ADHD," *Int. J. Spec. Educ.*, vol. 39, no. 1, pp. 44–52, 2024, doi: 10.52291/ijse.2024.39.5.

[24] S. Lorente-Echeverría, A. Corral-Abós, I. C. Lacruz, and B. M. Pardo, "Teachers' Profile in Sustainability: Association With Personal and Social Responsibility in Physical Education Classes," *J. Phys. Educ.*, vol. 34, no. 1, pp. 1–11, 2024, doi: 10.4025/jphyseduc.v34i1.3459.

[25] S. Ren, "Integration and Optimization of Educational Teaching Resources in Colleges and Universities Based on Clustering Algorithm," *J. Electr. Syst.*, vol. 20, no. 6s, pp. 2156–2165, 2024, doi: 10.52783/jes.3130.

[26] J. A. Ligibel, Y. Zheng, W. T. Barry, T. Sella, K. J. Ruddy, M. L. Greany, S. M. Resenberg, K. M. Emmons, and A. H. Patridge, "Effects of an Educational Physical Activity Intervention in Young Women With Newly Diagnosed Breast Cancer: Findings From the Young and Strong Study," *Cancer*, vol. 129, no. 14, pp. 2135–2143, 2023, doi: 10.1002/cncr.34779.

[27] N. Hwang, P. Graff, and M. Berends, "Timing and Frequency Matter: Same Race/Ethnicity Teacher and Student Achievement by School Level and Classroom Organization," *Educ. Policy*, vol. 37, no. 5, pp. 1349–1379, 2022, doi: 10.1177/08959048221087212.

[28] S. Glock, A. Shevchuk, C. Fuhrmann, and S. Rahn, "Role of Gender Match Between Students and Teachers and Students' Ethnicity in Teacher–student Relationships," *Learn. Environ. Res.*, vol. 27, no. 3, pp. 745–760, 2024, doi: 10.1007/s10984-024-09499-9.

[29] N. Gul, M. Bibi, M. Bilal, S. Ilyas, N. Saba, and I. H. Khan, "Factors Affecting Academic Performance of BS Students in Mansehra City (Pakistan)," *J. Asian Dev. Stud.*, vol. 13, no. 1, pp. 329–340, 2024, doi: 10.62345/jads.2024.13.1.28.

[30] L. Segkulu, "Factors Affecting Students' Academic Performance in Social Studies Subject - The Case of Selected Senior High Schools in Sangnarigu District and Tamale Metropolis of Northern Ghana," *Asian J. Educ. Soc. Stud.*, vol. 34, no. 2, pp. 57–65, 2022, doi: 10.9734/ajess/2022/v34i2748.

[31] M. M. Abubakar, I. S. Jummai, S. I. Danjuma, B. J. Nadikko, and A. M. Yusuf, “Psychological Factors as Correlates of Undergraduates Students’ Academic Performance in Educational Psychology, Gombe State University,” *Integr. J. Educ. Train.*, vol. 6, no. 1, pp. 8–11, 2022, doi: 10.31248/ijet2021.127.

[32] M. Faisal, N. Nurdin, F. Fajriana, and Z. Fitri, “Information and Communication Technology Competencies Clustering for Students for Vocational High School Students Using K-Means Clustering Algorithm,” *Int. J. Eng. Sci. Inf. Technol.*, vol. 2, no. 3, pp. 111–120, 2022, doi: 10.52088/ijesty.v2i3.318.

[33] W. Li, “Factors Influencing Student Engagement and Behavioral Differences Based on K-Means Cluster Analysis,” *J. Comput. Methods Sci. Eng.*, vol. 25, no. 2, pp. 1835–1844, 2024, doi: 10.1177/14727978241305758.

[34] R. Liu, “Data Analysis of Educational Evaluation Using K-Means Clustering Method,” *Comput. Intell. Neurosci.*, vol. 2022, no. 1, pp. 1–10, 2022, doi: 10.1155/2022/3762431.

[35] E. Cahapin, B. Malabag, C. S. Santiago, J. L. Reyes, G. S. Legaspi, and K. L. Adrales, “Clustering of Students Admission Data Using K-Means, Hierarchical, and DBSCAN Algorithms,” *Bull. Electr. Eng. Informatics*, vol. 12, no. 6, pp. 3647–3656, 2023, doi: 10.11591/eei.v12i6.4849.