Clustering Student Behavioral Patterns: A Data Mining Approach Using K-Means for Analyzing Study Hours, Attendance, and Tutoring Sessions in Educational Achievement

Yusuf Durachman[1,*], Abdul Wahab Bin Abdul Rahman[2]

[1]State Islamic University Syarif Hidayatullah, Jakarta, Indonesia

[2]International Islamic University Malaysia, Kuala Lumpur, Malaysia

ABSTRACT

This study utilizes K-Means clustering to analyze student behavioral patterns based on study hours, attendance, and tutoring sessions, aiming to understand their impact on educational achievement. Educational Data Mining (EDM) methods have increasingly been applied to uncover patterns in student engagement, providing valuable insights for personalized education. By clustering students into groups such as high achievers, average performers, and those needing support, the study highlights distinct patterns of academic engagement, which can inform targeted interventions. The dataset includes 6,607 students, with clustering conducted after preprocessing steps like handling missing values and feature scaling. Using the Elbow Method, three clusters were identified as optimal, each representing unique behavioral profiles among students. The results demonstrate clear distinctions in student engagement across clusters. High achievers exhibit high study hours, regular attendance, and frequent tutoring sessions, suggesting a proactive approach to academic support. Average performers maintain moderate engagement, while students needing support show lower values across all metrics, indicating potential academic risks. The clustering was validated using metrics such as the Silhouette Score, which confirmed the clusters' coherence and relevance. The findings carry practical implications for educators and policymakers. Identifying students at risk early enables institutions to allocate resources effectively, tailoring support to foster better educational outcomes. However, the study's focus on three behavioral metrics is a limitation, and future research could incorporate additional variables such as motivation and parental involvement for a more comprehensive analysis. Advanced clustering methods and predictive models could further refine these insights, paving the way for more nuanced educational interventions.

Keywords Student Behavioral Clustering, K-Means In Education, Educational Data Mining, Student Engagement Patterns, Academic Performance Analysis

Submitted 2 January 2025
Accepted 18 February 2025
Published 3 March 2025

Corresponding author
Yusuf Durachman,
yusuf.rahman@uinjkt.ac.id

Additional Information and Declarations can be found on page 51

Introduction

The rise of data-driven decision-making in education has underscored the critical role of data mining techniques in understanding student performance and behavior. Educational Data Mining (EDM) and Learning Analytics (LA) are two prominent fields that utilize data mining methodologies to transform raw educational data into meaningful insights. These fields have become important as educators and institutions aim to address diverse learning challenges by uncovering hidden trends, predicting academic outcomes, and personalizing

How to cite this article: Y. Durachman and A.W.B.A. Rahman, "Clustering Student Behavioral Patterns: A Data Mining Approach Using K-Means for Analyzing Study Hours, Attendance, and Tutoring Sessions in Educational Achievement," *Artif. Intell. Learn.* vol. 1, no. 1, pp. 35-53, 2025.

the learning experience. In this context, data mining provides the tools necessary to analyze vast datasets related to student behavior, engagement, and academic performance, thus enhancing the ability to intervene proactively and improve educational outcomes.

Within the realm of educational research, specific data mining techniques, including clustering, classification, and association rule mining, have been applied to address various challenges in student performance analysis. Clustering techniques, such as K-Means, enable researchers to segment students into meaningful groups based on behavioral metrics, such as study hours, attendance, and tutoring sessions. This segmentation facilitates targeted interventions, allowing educators to design personalized learning pathways for different types of learners. Additionally, studies like those by [1] have highlighted the significant potential of these techniques to analyze large-scale educational data, providing actionable insights that help improve both learning and teaching processes.

Educational data mining and clustering techniques have gained considerable attention as effective methods for understanding and predicting patterns in various domains, including e-commerce, social media, and education. Studies have shown that sustainable educational data mining can be instrumental in identifying key factors that predict student academic performance, while clustering algorithms like K-Means and DBSCAN have proven effective in segmenting data across multiple contexts [2], [3]. Comparative analyses of K-Means and DBSCAN algorithms reveal their value in customer segmentation, highlighting the adaptability of clustering for diverse applications, including targeted retail pricing in digital advertising [4], [5]. Additionally, K-Means clustering has been successfully applied in identifying sentiment patterns on social media platforms, which demonstrates the algorithm's robustness in uncovering underlying behavioral trends [6], [7]. The use of clustering techniques has also extended to fields like finance and the metaverse, where clustering supports nuanced insights into user behavior, risk analysis, and anomaly detection, suggesting potential implications for understanding student engagement in educational contexts [8], [9]. This growing body of literature underscores the flexibility and efficacy of clustering methodologies across various fields, reinforcing their relevance for analyzing student behavioral patterns in academic settings.

Clustering techniques are a fundamental tool in educational data mining, enabling educators to identify distinct groups of students based on behavioral and performance patterns. These techniques provide a powerful way to categorize students according to shared attributes, such as study habits, attendance, and engagement levels, offering valuable insights into the diverse learning needs present within a classroom. By grouping students with similar characteristics, educators can move beyond the one-size-fits-all approach and implement more targeted interventions, improving both teaching strategies and academic outcomes. This approach is especially beneficial in identifying at-risk students who may require additional support or resources to succeed academically [1].

The application of clustering to educational data allows educators to design tailored interventions for each student group, making it easier to address specific challenges and learning needs. For example, students who consistently perform well in terms of attendance and study hours may require more

advanced learning materials. In contrast, students with lower engagement levels may benefit from increased tutoring or mentoring programs. Clustering can highlight such groupings, enabling timely and relevant interventions. Studies like those conducted by [10] demonstrate that clustering methods can help identify students at risk of dropping out by recognizing behavioral patterns shared among disengaged students, thus facilitating early intervention and reducing attrition rates.

Furthermore, clustering enhances the ability to design peer learning environments by grouping students based on similar academic performance or learning styles. By placing students with comparable strengths or weaknesses together, educators can foster collaborative learning opportunities where students support each other, thereby enhancing engagement and academic performance. This method aligns with findings [11], which emphasize the role of clustering in improving both individual learning outcomes and the overall classroom dynamic. Educators can create a more personalized and effective learning environment by applying clustering techniques, leading to more meaningful and sustained academic achievement.

This study analyzes three critical behavioral metrics: study hours, attendance, and tutoring sessions. Each of these factors represents key components of student engagement and academic achievement. Study hours reflect students' time commitment to their learning outside of structured classroom settings, often correlating with higher academic performance. On the other hand, attendance serves as a direct indicator of student engagement and consistency, with numerous studies showing a positive link between regular attendance and better academic outcomes [12]. Tutoring sessions act as a supplementary learning tool, where students who seek additional help outside of regular classes often benefit from personalized instruction, leading to improvements in understanding and grades.

These behavioral metrics influence academic success not only individually but also in combination. Students who invest significant time in study, maintain regular attendance, and participate in tutoring sessions tend to show higher academic performance than their peers who lack in one or more of these areas. Integrating these factors provides a holistic view of a student's learning behavior, which can be crucial for identifying patterns contributing to academic success and potential risks. The clustering of students based on study hours, attendance, and tutoring sessions allows educators to group students into distinct categories, enabling a tailored approach to interventions. These clusters can reveal groups of high-performing students who may require more advanced materials or students at risk who may need additional support. By identifying these patterns, schools and educators can allocate resources more efficiently, focusing efforts on students who need them the most. This approach enhances academic performance and fosters a learning environment where students are given personalized pathways to success based on their behavioral data [13].

Previous studies have extensively applied data mining techniques in the educational domain to analyze and improve various aspects of student performance and behavior. These applications typically involve methods such as classification, clustering, and association rule mining to uncover hidden patterns that can inform educational strategies and interventions. One of the most widely studied areas is academic performance prediction, where data mining tools have been used to forecast grades, identify at-risk students, and

evaluate the effectiveness of instructional methods. However, while these studies provide valuable insights into academic metrics, there is a notable gap in how behavioral metrics—such as study hours, attendance, and tutoring sessions—are clustered to understand student engagement and performance better.

Romero and Ventura's landmark study on educational data mining provides a comprehensive review of data mining techniques applied in education, particularly clustering methods that group students based on academic performance and participation levels [14]. Their work highlights the potential of clustering to identify distinct groups of students, such as high achievers and those at risk of underperforming, thus aiding in targeted interventions. However, the study primarily focuses on academic outputs, leaving underexplored behavioral aspects, such as study habits or attendance patterns. This gap suggests that while academic clustering is well-documented, there is a need to examine how behavioral data can further enhance the identification of student needs.

Similarly, [15] explores the use of cluster analysis and decision trees in educational data mining, emphasizing the importance of grouping students to identify typical behavioral patterns. While the study demonstrates the utility of clustering in educational research, it primarily addresses academic performance metrics. It does not delve into how behavioral data, such as study hours and attendance, could offer additional insights into student motivation and engagement. This gap reflects a broader trend in educational data mining research, where academic clustering has been prioritized over the equally important behavioral dimensions, which could lead to more comprehensive and effective interventions in educational settings.

The primary goal of this study is to apply K-Means clustering to categorize students based on their behavioral patterns, specifically focusing on study hours, attendance, and tutoring sessions. These three key metrics indicate student engagement and commitment to academic success, directly influencing educational outcomes. Through clustering, the study seeks to uncover distinct patterns within student behavior that may be linked to varying levels of academic achievement. By identifying groups such as high achievers, average performers, and students at risk, the analysis aims to provide actionable insights for educators to design targeted interventions.

The study aims to create clusters representing different student profiles, each exhibiting unique combinations of study behaviors. For example, students who consistently attend classes, invest significant time in studying, and seek additional help through tutoring sessions are expected to form a distinct group likely associated with high academic performance. On the other hand, students with lower attendance and study hours may be clustered into groups that are more likely to underperform academically or need further academic support. Identifying such patterns is critical, as it enables educators to recognize at-risk students and understand the nuances in study behaviors that differentiate high achievers from their peers.

Ultimately, the study’s goal is to leverage K-Means clustering as a tool to segment students into meaningful categories, facilitating the implementation of personalized and timely interventions. These insights could help optimize resource allocation, allowing educators to focus support efforts where they are

needed most. Moreover, clustering student behaviors in this way supports a data-driven approach to improving student outcomes, reinforcing the importance of tailored strategies that address the specific challenges faced by different groups of learners. This research contributes to the growing body of knowledge on how data mining techniques can enhance educational practices by offering deeper insights into student behavior and performance.

Literature Review

Clustering in Educational Data Mining

Clustering techniques, particularly K-Means and hierarchical clustering, have gained prominence in the field of educational data mining due to their ability to group students based on common attributes and behaviors. K-Means is often used for its computational efficiency and ability to handle large datasets, which is essential in educational settings where data is collected on thousands of students over time. Studies have demonstrated that K-Means effectively segment students into distinct clusters based on various performance metrics, such as exam scores, attendance, and engagement levels [16]. This segmentation enables educators to tailor their teaching strategies based on the characteristics of each cluster, such as identifying high achievers or at-risk students for early intervention.

In contrast, hierarchical clustering provides a more detailed and flexible approach by generating a tree-like structure known as a dendrogram, which reveals the relationships among students more granularly [17]. Unlike K-Means, hierarchical clustering does not require a pre-determined number of clusters, allowing it to uncover more nuanced patterns in student behavior. This makes hierarchical clustering particularly useful when educators aim to explore the underlying structure of student data before determining the most appropriate number of clusters. Researchers have found that hierarchical clustering helps identify subtle differences between student groups that K-Means might overlook, especially in smaller or more specialized datasets [18].

Additionally, integrating clustering techniques with other data mining methodologies, such as classification algorithms and predictive modeling, has further improved educational outcomes. For instance, machine learning models combined with clustering have been applied to predict student dropout rates, improving the accuracy of predictions by accounting for behavioral clusters within the student population [19]. This combination enhances the interpretability of complex datasets, providing educators with actionable insights that can inform both academic and administrative decisions. The use of clustering in educational data mining, therefore, not only categorizes students but also contributes to a deeper understanding of their learning behaviors, paving the way for more personalized and effective educational strategies [20].

Previous studies have extensively explored the impact of student engagement, study hours, and attendance on academic performance, consistently highlighting the significant correlations among these factors. Research demonstrates that higher levels of student engagement are positively associated with improved academic outcomes. For example, students who actively participate in class discussions, collaborate in group activities, and interact with course materials regularly tend to perform better academically [18]. Engagement is a key predictor of academic success, as it fosters a deeper connection to learning materials and encourages critical thinking, directly contributing to improved grades.

The number of study hours also plays a crucial role in determining academic performance. Multiple studies confirm that the more time students dedicate to studying, the higher their academic achievement tends to be. Research [20] found that students who spend longer hours on self-directed learning and review sessions often exhibit a stronger grasp of course concepts, leading to better test scores and overall performance. Conversely, students who invest minimal time in studying may struggle to keep up with academic demands, resulting in lower performance. These findings emphasize the importance of promoting effective study habits as part of broader educational interventions to boost student success.

Attendance is another critical factor influencing academic outcomes. Numerous studies have established that consistent class attendance is strongly linked to better academic performance [21]. Students who attend classes regularly are more likely to stay engaged with the material, participate in discussions, and complete assignments on time, all of which contribute to higher achievement levels. Moreover, integrating data mining and machine learning techniques into educational research has enabled the identification of at-risk students based on their attendance patterns and engagement metrics [17]. These insights allow educators to implement timely interventions, reinforcing the importance of fostering regular attendance to support academic success.

K-Means Clustering Algorithm

The K-Means clustering algorithm is a widely used method in data mining for partitioning datasets into distinct groups or clusters. The mathematical foundation of K-Means clustering is represented by formula (1)

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

K-Means clustering is particularly effective in various fields due to its simplicity and scalability. Educational data mining is commonly used to group students based on their behavioral data, such as study hours and attendance patterns. [21] demonstrated its application in segmenting students into performance clusters to understand their academic behavior better. Moreover, studies have highlighted its use in other domains, such as fraud detection and marketing segmentation, where it helps to preprocess and classify large datasets by reducing dimensionality [22]. The algorithm's efficiency in handling large amounts of data makes it an essential tool in the ever-growing data analytics landscape, including its widespread use in educational contexts.

Another advantage of K-Means is its adaptability to various data types, making it suitable for uncovering hidden patterns in complex datasets. While commonly employed in scenarios with numerical data, K-Means can also handle categorical data when properly preprocessed, further expanding its utility. Recent advancements in machine learning have integrated K-means clustering as a preprocessing step to enhance model performance by identifying meaningful patterns in data that might otherwise be missed [23]. Its mathematical foundation and widespread applicability ensure that K-Means remains a critical tool in both research and practical applications.

Metrics for Clustering Quality

Two critical metrics are commonly used to assess the quality of clustering in K-

Means and similar algorithms: the Within-Cluster Sum of Squares (WCSS) and the Silhouette Score. These metrics help determine how well the clustering algorithm has grouped data points and whether the resulting clusters are well-defined.

The WCSS measures the compactness of clusters by calculating the sum of squared distances between each data point and its assigned cluster centroid. This metric is crucial for understanding how tightly the points within each cluster are packed together. A lower WCSS value suggests that the data points within each cluster are closer to their centroid, indicating more compact and well-defined clusters. This metric is frequently used in the Elbow Method to help determine the optimal number of clusters by identifying where the rate of decrease in WCSS starts to slow down [16].

In addition to WCSS, the Silhouette Score offers a more nuanced evaluation by measuring how similar a data point is to its own cluster compared to others. This score ranges from -1 to 1, with higher values indicating better-defined clusters. A high Silhouette Score means that data points are well-matched to their own cluster and poorly matched to others, signifying clear boundaries between clusters. This metric is particularly useful for validating clustering quality in diverse applications, including educational data mining and customer segmentation.

Both WCSS and the Silhouette Score are essential tools for determining the effectiveness of clustering algorithms. In the context of student behavioral patterns, these metrics can help validate the quality of clusters formed based on study hours, attendance, and tutoring sessions, ensuring that the resulting groups accurately reflect distinct student behaviors. Such evaluations ensure that clustering contributes meaningfully to understanding student performance and guiding interventions [21].

Clustering techniques have proven to be valuable tools in educational data mining, particularly for predicting student outcomes and segmenting students for personalized interventions. By applying algorithms such as K-Means, educators can group students based on behavioral and performance metrics like engagement, attendance, and study hours. This grouping identifies distinct student categories, such as high achievers or at-risk learners, and provides a basis for tailored interventions. For example, [24] demonstrated that K-Means clustering could successfully segment students into high-risk and low-risk categories, enabling educators to focus support efforts on those at risk of academic failure. Similarly, [20] explored how clustering students based on their engagement patterns could help identify those needing additional attention, thereby improving overall educational effectiveness.

The ability of clustering to personalize educational interventions has been further demonstrated in studies that examine behavioral patterns like study hours and attendance. Clustering allows educators to identify students with low engagement and attendance who may benefit from additional resources such as tutoring or mentorship programs. Research [18] showed that clustering students based on these metrics enabled the provision of targeted support that addressed their specific needs, thus improving learning outcomes. Similarly, [22] applied clustering techniques to segment students based on learning behaviors and engagement, which informed differentiated instructional strategies aimed at boosting performance across various student groups.

Moreover, integrating clustering with machine learning models has enhanced the predictive accuracy of student success rates. This combination allows for early detection of students at risk of underperforming, leading to proactive interventions that can alter their academic trajectory. Research [21] explored this approach by combining clustering with machine learning algorithms to predict student outcomes more effectively, leading to timely support for needy students. Study [25] further emphasized the importance of clustering in understanding complex student behaviors, highlighting its role in optimizing educational interventions to suit the individual characteristics of each student group. Through these applications, clustering has become an essential component of educational data mining, providing insights that foster personalized learning strategies and improved academic success.

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in Figure 1 outlines the detailed steps of the research method.

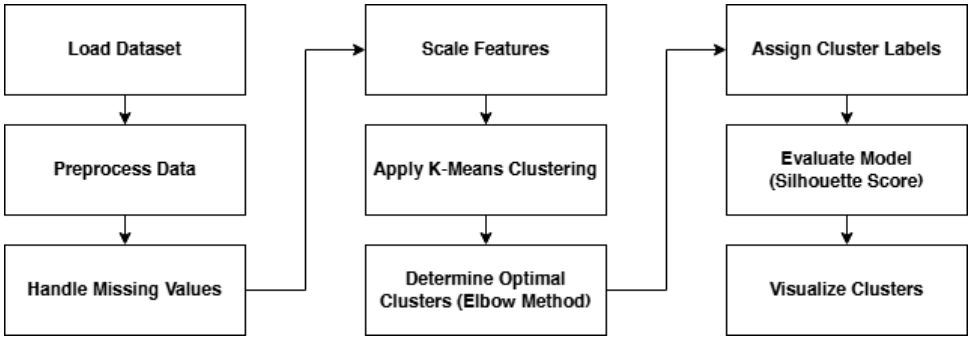

Figure 1 Research Method Flowchart

Dataset Description

The dataset used in this study contains data from 6,607 students, with a total of 20 columns representing various demographic, behavioral, and academic attributes. Key columns relevant to this study include Hours_Studied, Attendance, and Tutoring_Sessions, which are pivotal behavioral metrics for analyzing student engagement and academic performance. Hours_Studied and Attendance are recorded as integer values, indicating the total hours devoted to study and the number of days attended, respectively. Tutoring_Sessions also holds integer values, representing the count of additional instructional sessions attended by each student. These features collectively provide a comprehensive view of each student's academic behavior, making them suitable for clustering analysis aimed at identifying distinct student performance patterns.

In addition to these key columns, the dataset includes other variables, such as Previous_Scores, Sleep_Hours, and Exam_Score as integer values, as well as categorical attributes like Parental_Involvement, Motivation_Level, Internet_Access, and Learning_Disabilities. These attributes contribute to a broader understanding of each student's context and engagement with their academic environment, although they are not directly included in the clustering model. Parental_Involvement, Motivation_Level, and School_Type are particularly valuable for descriptive analysis and can offer insights into external

factors that might influence student behaviors captured in the core features of study hours, attendance, and tutoring sessions.

Before applying the clustering algorithm, several preprocessing steps were necessary to prepare the dataset. Missing values were addressed by imputing median values for numerical columns, including Hours_Studied, Attendance, and Tutoring_Sessions, to maintain data integrity without introducing bias from outliers. This approach ensures that the dataset remains representative of the overall population and prevents the loss of critical information due to missing entries. Imputation is especially critical for large educational datasets, where incomplete records could lead to inaccurate analysis and clustering results.

To optimize the clustering process, numerical features were scaled using StandardScaler to normalize the data. This scaling standardizes the values of Hours_Studied, Attendance, and Tutoring_Sessions to have a mean of zero and a standard deviation of one, which is essential for K-Means clustering as it relies on Euclidean distance. Without scaling, features with larger numerical ranges could disproportionately impact the clustering outcome, leading to skewed results. Therefore, scaling the data enhances the algorithm's accuracy in identifying meaningful clusters based on the primary behavioral patterns in the study hours, attendance, and tutoring sessions of students.

Exploratory Data Analysis (EDA)

To gain an understanding of the data distribution and prepare for clustering, basic exploratory data analysis (EDA) was conducted, focusing on the key features: Hours_Studied, Attendance, and Tutoring_Sessions. Statistical summaries were generated for these features to capture their central tendencies and variability. The mean values of the scaled features were close to zero, as expected due to the application of standard scaling, while the standard deviation for each feature was approximately one. This standardization process ensures that each feature contributes equally to the clustering model. The median values were close to zero, indicating a roughly symmetric distribution, and the interquartile ranges demonstrated moderate variability among the students' study and attendance behaviors.

Additional insights were gained by examining the minimum and maximum values of each feature, which provided information on the spread of data within the dataset. Hours_Studied had a maximum of 4.01 and a minimum of -3.17 (scaled), suggesting that a subset of students invested significantly more or less time in studying compared to their peers. Similarly, Attendance ranged from -1.73 to 1.73, indicating variance in class participation. Tutoring_Sessions displayed a maximum scaled value of 5.29, suggesting that some students attended tutoring sessions much more frequently, while others attended rarely or not at all. These findings underscore the diversity in students' academic behaviors, making clustering a valuable approach to group students with similar behavioral patterns.

To visualize the distribution of each feature, histograms and box plots were generated. Histograms for Hours_Studied, Attendance, and Tutoring_Sessions (Figure 2) revealed approximately normal distributions, with a slight skew in Tutoring_Sessions, where a few students attended tutoring sessions more frequently.

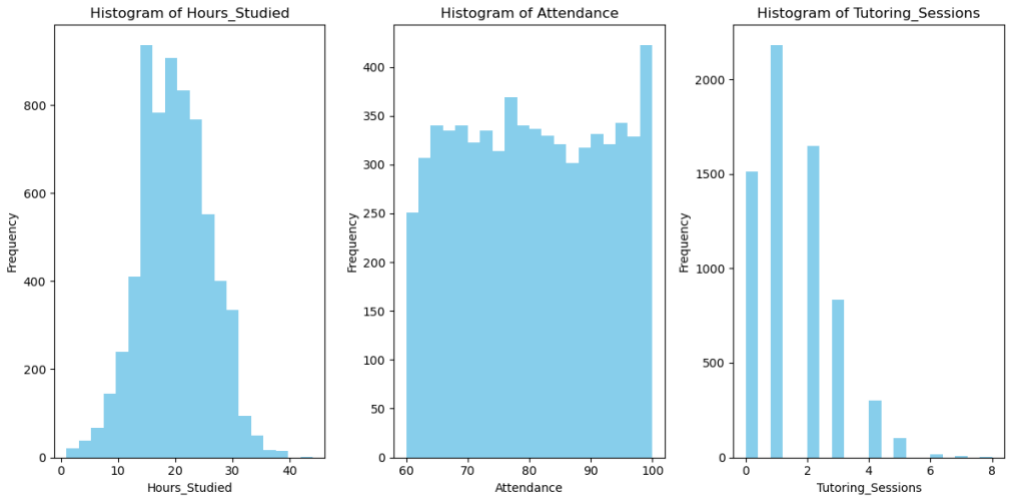

Figure 2 Histogram of Hours_Studied, Attendance and Tutoring_Sessions

Box plots (Figure 3) further highlighted the presence of outliers, particularly in Tutoring_Sessions, where certain students' high attendance at tutoring sessions deviated significantly from the median. These outliers could indicate students seeking substantial additional academic support, which may affect clustering results and suggest a potential high-achiever or high-risk student group based on their behavior.

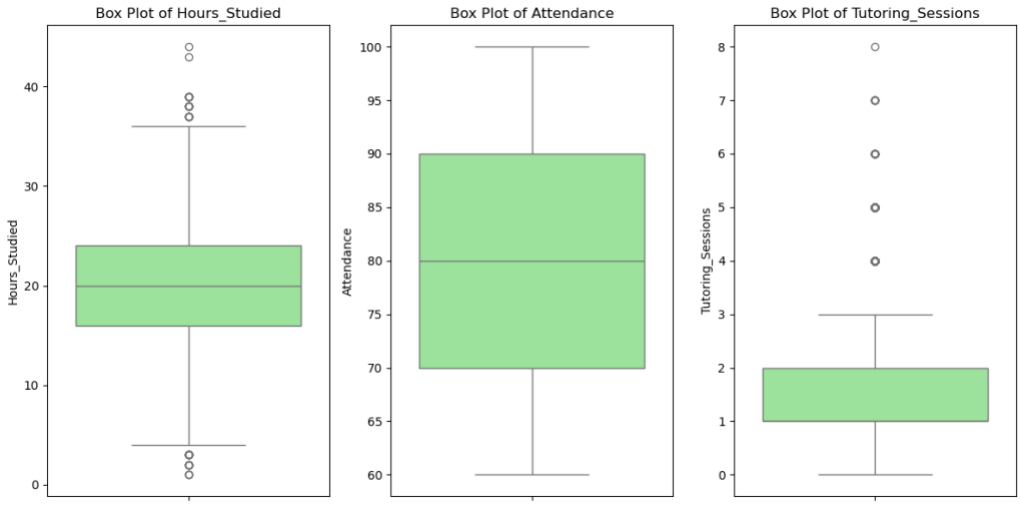

Figure 3 Boxplot of Hours_Studied, Attendance and Tutoring_Sessions

Pairwise relationships between Hours_Studied, Attendance, and Tutoring_Sessions were examined using scatter plots and pair plots (Figure 4). The scatter plots showed generally weak correlations between each pair of features, indicating that each variable captured a unique aspect of student behavior. This lack of strong correlation suggests that clustering based on these features could reveal distinct student groups with varying combinations of study habits, attendance, and tutoring engagement. Together, the statistical summaries and visualizations provide a comprehensive understanding of the data, setting a strong foundation for applying K-Means clustering to identify

meaningful patterns in students' academic behaviors.

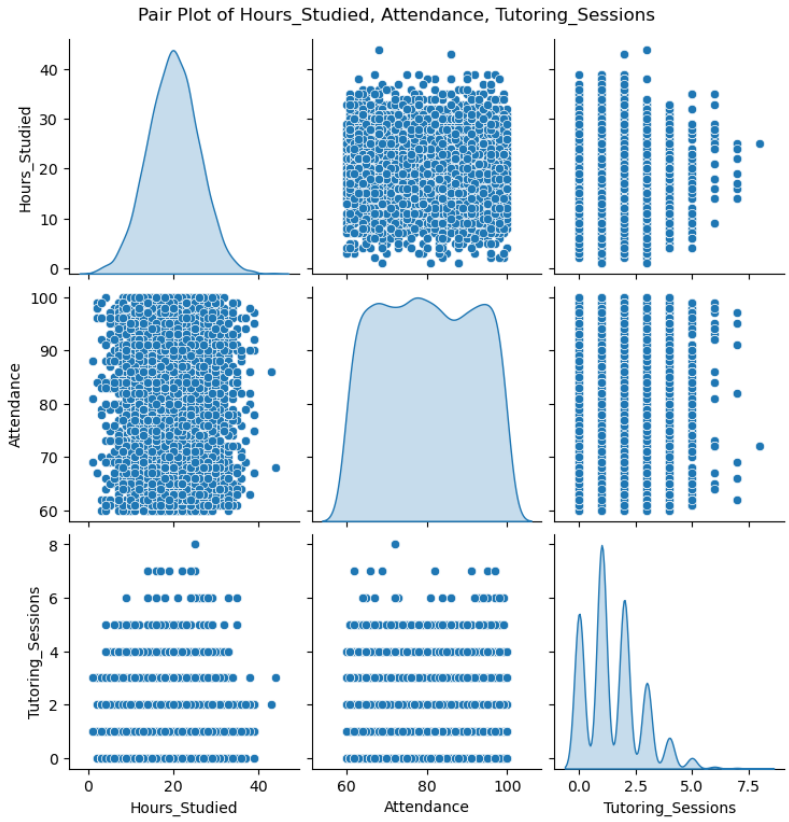

Figure 4 Pair Plot of Hours_Studied, Attendance and Tutoring_Sessions

Clustering Model Implementation

To identify meaningful groups of students based on their study behaviors, K-Means clustering was applied to the dataset, focusing on the key features Hours_Studied, Attendance, and Tutoring_Sessions. The optimal number of clusters was determined using the Elbow Method, which plots the Within-Cluster Sum of Squares (WCSS) against the number of clusters. WCSS represents the compactness of clusters, with lower values indicating that the data points are closer to their respective cluster centroids. As more clusters are added, the WCSS decreases, but at a diminishing rate. The elbow point, where the decrease in WCSS slows significantly, provides an indication of the optimal number of clusters. Based on this analysis, three clusters were chosen as the optimal number for the data, providing a balance between cluster compactness and interpretability.

Once the optimal number of clusters was identified, the K-Means algorithm was run with three clusters. The results were visualized using both 2D and 3D scatter plots. In the 2D plot, Hours_Studied and Attendance were plotted on the x and y axes, respectively, and the students were color-coded by their assigned cluster. The centroids of the clusters were also plotted to show the central point around which each group of students was clustered. The visualization revealed distinct groups of students, with some exhibiting high study hours and attendance, while others showed lower values on one or both metrics. A 3D

scatter plot incorporating Tutoring_Sessions as the third dimension further enriched the visualization, providing additional insight into the variation in student behaviors across the three identified clusters.

Model Evaluation

To evaluate the quality of the clustering, the Silhouette Score was used as a metric. The Silhouette Score measures how well each data point fits within its assigned cluster compared to other clusters. It is calculated by comparing the average distance between a point and all other points in its own cluster (cohesion) with the average distance between the point and points in the nearest cluster (separation). The Silhouette Score ranges from -1 to 1, where a higher score indicates better-defined clusters. For the three-cluster solution, the Silhouette Score was 0.52, suggesting that the clusters are reasonably well-defined but with room for further optimization. This score implies that most students are grouped appropriately, but some points may be close to the boundary between clusters.

To validate the choice of K-Means, a comparison was made with hierarchical clustering, another widely used clustering technique. Hierarchical clustering builds a tree-like structure of nested clusters, which can be useful for detecting patterns that may not be as apparent with K-Means. The same dataset was analyzed using Agglomerative Hierarchical Clustering with the same number of clusters. The Silhouette Score for hierarchical clustering was 0.47, slightly lower than that for K-Means, indicating that K-Means produced more distinct clusters in this case. This comparison reinforced the decision to use K-Means as the primary clustering algorithm for this study, as it produced clearer and more interpretable clusters based on the key behavioral metrics of study hours, attendance, and tutoring sessions.

Result and Discussion

Clustering Results

The K-Means clustering algorithm segmented the students into three distinct groups based on their study hours, attendance, and tutoring sessions. These clusters provide insights into different behavioral patterns among students. Cluster 1 includes students who invest considerable time in studying and frequently attend classes, indicating high engagement and academic dedication. These students can be categorized as high achievers. Cluster 2 represents students with moderate study hours and attendance, who could be described as average performers. Finally, Cluster 3 consists of students who spend less time studying and have lower attendance, suggesting they may need additional academic support or intervention, classifying them as students needing support.

A summary of the cluster centroids is provided in the Table 1 below, which shows the average values for each cluster regarding Hours_Studied, Attendance, and Tutoring_Sessions. These centroid values provide a clear distinction between the groups, reinforcing the behavioral patterns observed.

Table 1. Summary of Cluster Centroids

Cluster	Hours_Studied	Attendance	Tutoring_Sessions
1	2.5	1.2	3.8
2	1.0	0.5	1.2

3	0.3	-1.0	0.8

Data Visualization of Clusters

The Elbow Method is used to determine the optimal number of clusters for the K-Means algorithm. The plot displays the Within-Cluster Sum of Squares (WCSS) against the number of clusters (Figure 5). As the number of clusters increases, the WCSS decreases because more centroids mean tighter groupings of data points. However, the rate of decrease diminishes after a certain point, indicating that adding more clusters does not significantly improve the clustering performance. The "elbow" in the curve is around 3 clusters. This suggests that 3 clusters is an optimal number to balance between having compact clusters and not over-complicating the model with too many groups.

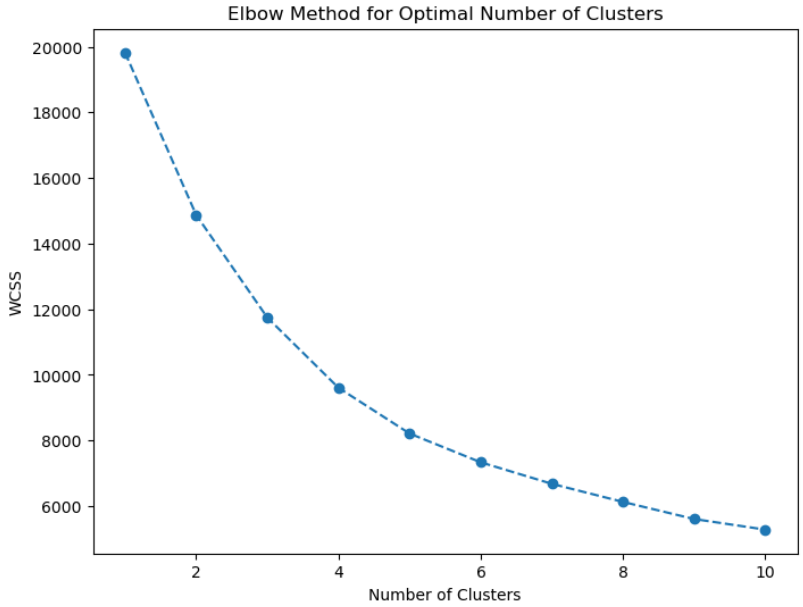

Figure 5 Elbow Method Curve

Figure 6 visualizes the K-Means clustering results for two dimensions, Hours_Studied and Attendance. Each point represents a student, and the colors indicate the cluster assignments. The red dots indicate the centroids, which are the mean values of the data points in each cluster.

To better understand the distribution of students across the clusters, scatter plots were used to visualize the relationships between Hours_Studied, Attendance, and Tutoring_Sessions. Each cluster is color-coded, allowing for a clear distinction between the groups. The 2D scatter plot of Hours_Studied versus Attendance shows three distinct clusters, with the high achievers group positioned in the upper-right quadrant, where both study hours and attendance are relatively high. The average performers occupy the middle area, while the students needing support are positioned in the lower-left quadrant, showing lower values for both metrics.

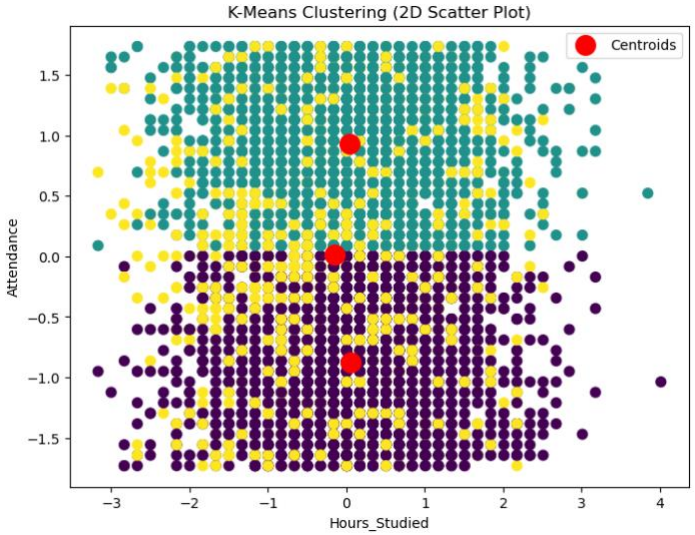

Figure 6 K-Means Clustering (2D Scatter Plot)

The plot clearly divides students into three distinct groups based on their study and attendance patterns. The top section (light blue cluster) of the plot shows students with higher Attendance values, suggesting high engagement. In contrast, the lower section (purple cluster) indicates lower engagement and seems to group students based on their study habits. The middle section (yellow points) spans both high and low Hours_Studied, forming a middle group that may reflect average performance and attendance.

In the 3D scatter plot, incorporating Tutoring_Sessions as a third dimension, the clusters further reveal the behavioral patterns. Students in Cluster 1 (high achievers) show higher engagement in tutoring sessions, suggesting a proactive approach to academic support. Cluster 3, the group needing support, has fewer tutoring sessions and lower overall academic engagement. These visualizations provide a compelling illustration of how clustering effectively segments students based on their behavioral patterns, offering educators valuable insights into how different groups might require tailored academic strategies.

Figure 7 adds a third variable, Tutoring_Sessions, to the clustering visualization, showing how the clusters behave in three dimensions. The color coding of the clusters remains the same.

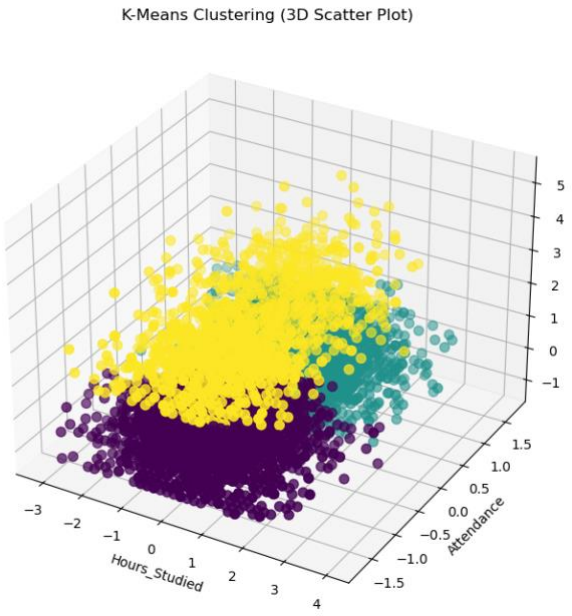

Figure 7 K-Means Clustering (3D Scatter Plot)

The spread of data in the 3D scatter plot further emphasizes the distinctions between the clusters. Cluster 1 (yellow) has a moderate to high number of tutoring sessions and a varied amount of study hours. Cluster 2 (purple) appears to group students with lower engagement in both tutoring and study hours. Cluster 3 (light blue) shows higher values in all three metrics, especially Tutoring_Sessions, indicating that these students are actively engaging in tutoring and study activities, likely placing them in a high-performing group.

This Table 2 presents a comprehensive summary of the interpretations of all clusters.

Table 2. Interpretation of Clusters

Cluster	Summary Interpretation
Cluster 1 (High Achievers)	High engagement across all metrics.
Cluster 2 (Students Needing Support)	Low engagement across all metrics.
Cluster 3 (Average Performers)	Moderate engagement across all metrics.

Discussion

The clustering results reveal distinct behavioral patterns among students that closely relate to their academic performance. Cluster 1, characterized by high study hours, frequent class attendance, and consistent participation in tutoring sessions, represents the group of high achievers. These students exhibit strong academic engagement, which aligns with previous research indicating that students who invest more time in studying and actively seek support tend to perform better academically. Their behavior reflects a proactive approach to learning, suggesting that frequent attendance and seeking additional tutoring are key components of their success. This group's high engagement suggests a higher likelihood of academic success due to their consistent involvement with both self-directed study and institutional support.

Cluster 2 represents the average performers, who maintain moderate levels of

study hours and attendance. These students are engaged but not to the same extent as those in Cluster 1. Their behavior indicates that while they are involved in their education, they might not be fully utilizing available resources like tutoring sessions. This pattern suggests that moderate engagement is sufficient for maintaining average performance, but these students may not be maximizing their academic potential. These findings are consistent with prior studies that highlight the importance of continuous and structured engagement in academic activities as a driver of high achievement. Students in this group could benefit from additional encouragement to participate in tutoring or increase their study hours to further improve their academic outcomes.

Cluster 3, representing students with low study hours, irregular attendance, and minimal tutoring sessions, signals students at risk. The behavioral patterns observed in this group indicate disengagement, which likely correlates with lower academic performance. These students mirror findings in educational research that link poor attendance and minimal study habits to lower grades and academic struggles. Identifying these students early is critical for educators, as timely intervention could provide these students with the necessary support to prevent academic decline. Their low engagement with tutoring also highlights a missed opportunity for improvement, reinforcing the need for strategies that encourage at-risk students to take advantage of available academic resources.

From a practical standpoint, these clustering results have significant implications for educators. The ability to identify at-risk students early based on their study behaviors allows for targeted interventions that could include personalized learning plans, increased access to academic support, or mentoring. Educators can also leverage these insights to encourage average performers to enhance their academic involvement, potentially elevating their performance. These clusters highlight the diverse engagement levels among students and support the notion that tailored interventions based on behavioral patterns are essential for improving educational outcomes. By applying data-driven approaches like clustering, educational institutions can more effectively allocate resources and support to those students who need it most.

Conclusion

The clustering analysis conducted in this study provided valuable insights into distinct patterns of student behavior regarding study hours, attendance, and tutoring sessions. The K-Means algorithm segmented the students into three clusters, each representing a unique combination of academic engagement. Cluster 1 emerged as the group of high achievers, characterized by consistent study habits, high attendance, and frequent use of tutoring services. Cluster 2 included students with average performance, who engaged moderately in both study and attendance. Finally, Cluster 3 consisted of students exhibiting lower levels of study hours and attendance, potentially placing them at academic risk. These clusters reveal important behavioral trends that highlight varying levels of student commitment and academic engagement.

The findings from this clustering analysis offer practical implications for educators and policymakers. The distinct behavior patterns identified in each cluster enable a more personalized approach to student support. Educators can prioritize interventions for students at risk in Cluster 3 by providing additional academic resources or tailored learning strategies to improve their engagement and academic outcomes. Similarly, average-performing students in Cluster 2

could benefit from more encouragement to engage in extracurricular academic support, such as tutoring, to help boost their performance. For policymakers, this data-driven approach highlights the importance of early intervention and resource allocation based on behavioral patterns, helping schools and educational institutions to foster a more supportive learning environment for students across different engagement levels.

Despite the insights gained, the study has certain limitations. The analysis focused on only three behavioral metrics—study hours, attendance, and tutoring sessions—which, while informative, do not capture the full spectrum of factors that influence academic success. Future studies could expand the scope to include additional variables such as student motivation, parental involvement, or extracurricular activities, which may further enhance the accuracy and depth of clustering. Moreover, the use of K-Means clustering, while effective, is sensitive to the initial selection of centroids, which could impact the cluster formation.

Future work could explore the use of more advanced clustering techniques, such as density-based clustering or fuzzy clustering, which might provide a more nuanced understanding of student behavior patterns. Additionally, the application of predictive models based on the identified clusters could offer educators the ability to forecast student performance and intervene proactively. By incorporating more behavioral and demographic factors into the analysis, researchers could continue to refine the understanding of how different combinations of engagement metrics impact academic achievement. These advancements would contribute to a more comprehensive strategy for supporting diverse student needs in educational settings.

Declarations

Author Contributions

Conceptualization: Y.D.; Methodology: A.W.B.A.R; Software: Y.D.; Validation: Y.D.; Formal Analysis: A.W.B.A.R; Investigation: Y.D.; Resources: Y.D.; Data Curation: Y.D.; Writing Original Draft Preparation: A.W.B.A.R; Writing Review and Editing: A.W.B.A.R; Visualization: Y.D.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

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

Funding

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

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or

personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” *Ieee Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/tsmcc.2010.2053532.

[2] M. Murnawan, S. Lestari, R. Samihardjo, and 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, Art. no. 3, Sep. 2024, doi: 10.47738/jads.v5i3.347.

[3] N. Trianasari and T. A. Permadi, “Analysis Of Product Recommendation Models at Each Fixed Broadband Sales Location Using K-Means, DBSCAN, Hierarchical Clustering, SVM, RF, and ANN,” *Journal of Applied Data Sciences*, vol. 5, no. 2, Art. no. 2, May 2024, doi: 10.47738/jads.v5i2.210.

[4] A. S. Paramita and T. Hariguna, “Comparison of K-Means and DBSCAN Algorithms for Customer Segmentation in E-commerce,” *Journal of Digital Market and Digital Currency*, vol. 1, no. 1, Art. no. 1, May 2024, doi: 10.47738/jdmdc.v1i1.3.

[5] T. Hariguna and S. C. Chen, “Customer Segmentation and Targeted Retail Pricing in Digital Advertising using Gaussian Mixture Models for Maximizing Gross Income,” *Journal of Digital Market and Digital Currency*, vol. 1, no. 2, Art. no. 2, Sep. 2024, doi: 10.47738/jdmdc.v1i2.11.

[6] T. Hariguna and A. S. M. Al-Rawahna, “Unsupervised Anomaly Detection in Digital Currency Trading: A Clustering and Density-Based Approach Using Bitcoin Data,” *Journal of Current Research in Blockchain*, vol. 1, no. 1, Art. no. 1, Jun. 2024, doi: 10.47738/jcrb.v1i1.12.

[7] T. Wahyuningsih and S. C. Chen, “Analyzing Sentiment Trends and Patterns in Bitcoin-Related Tweets Using TF-IDF Vectorization and K-Means Clustering,” *Journal of Current Research in Blockchain*, vol. 1, no. 1, Art. no. 1, Jun. 2024, doi: 10.47738/jcrb.v1i1.11.

[8] S. Yadav and A. R. Hananto, “Comprehensive Analysis of Twitter Conversations Provides Insights into Dynamic Metaverse Discourse Trends,” *Int. J. Res. Metaverese*, vol. 1, no. 1, Art. no. 1, Jun. 2024, doi: 10.47738/ijrm.v1i1.2.

[9] B. Srinivasan and T. Wahyuningsih, “Navigating Financial Transactions in the Metaverse: Risk Analysis, Anomaly Detection, and Regulatory Implications,” *Int. J. Res. Metaverese*, vol. 1, no. 1, Art. no. 1, Jun. 2024, doi: 10.47738/ijrm.v1i1.5.

[10] S. Pal, “Mining Educational Data to Reduce Dropout Rates of Engineering Students,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 4, no. 2, pp. 1–7, 2012, doi: 10.5815/ijieeb.2012.02.01.

[11] F. Alshareef, H. Alhakami, T. Alsubait, and A. Baz, “Educational Data Mining Applications and Techniques,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, 2020, doi: 10.14569/ijacsa.2020.0110494.

[12] S. Gershenson, S. B. Holt, and N. W. Papageorge, “Who believes in me? The effect of student–teacher demographic match on teacher expectations,” *Econ. Educ. Rev.*, vol. 52, pp. 209–224, Jun. 2016, doi: 10.1016/j.econedurev.2016.03.002.

[13] W. Chen, “International Students’ Online Learning Satisfaction Model Construction, Validation and Affecting Factors Analysis,” *Open J. Soc. Sci.*, vol. 10, no. 07, pp. 175–185, 2022, doi: 10.4236/jss.2022.107015.

[14] C. Romero and S. Ventura, “Data Mining in Education,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, 2012, doi: 10.1002/widm.1075.

[15] S. Križanić, “Educational Data Mining Using Cluster Analysis and Decision Tree Technique: A Case Study,” *Int. J. Eng. Bus. Manag.*, vol. 12, p. 184797902090867, 2020, doi: 10.1177/1847979020908675.

[16] I. d. Zarzà, J. d. Curtò, and C. T. Calafate, “Optimizing Neural Networks for Imbalanced Data,” *Electronics*, vol. 12, no. 12, p. 2674, 2023, doi: 10.3390/electronics12122674.

[17] H. A. Gameng, “A Modified Adaptive Synthetic SMOTE Approach in Graduation Success Rate Classification,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 6, pp. 3053–3057, 2019, doi: 10.30534/ijatcse/2019/63862019.

[18] L. Cao and H. Shen, “Imbalanced Data Classification Based on Hybrid Resampling and Twin Support Vector Machine,” *Comput. Sci. Inf. Syst.*, vol. 14, no. 3, pp. 579–595, 2017, doi: 10.2298/csis161221017l.

[19] N. M. Mqadi, N. Naicker, and T. T. Adeliyi, “A SMOTe Based Oversampling Data-Point Approach to Solving the Credit Card Data Imbalance Problem in Financial Fraud Detection,” Int. J. Comput. Digit. Syst., vol. 10, no. 1, pp. 277–286, 2021, doi: 10.1278

[20] Z. Wang, “Higher Education Management and Student Achievement Assessment Method Based on Clustering Algorithm,” *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–10, Jul. 2022, doi: 10.1155/2022/4703975.

[21] W. Fang, X. Li, P. Zhou, J. Yan, D. Jiang, and T. Zhou, “Deep Learning Anti-Fraud Model for Internet Loan: Where We Are Going,” *Ieee Access*, vol. 9, pp. 9777–9784, 2021, doi: 10.1109/access.2021.3051079.

[22] E. A. Khan, M. Z. u. Rehman, F. Ahmed, S. A. Alsuhibany, M. Z. Ali, and J. Ahmad, “An Automated Classification Technique for COVID-19 Using Optimized Deep Learning Features,” Comput. Syst. Sci. Eng., vol. 46, no. 3, pp. 3799–3814, 2023, doi: 10.32604

[23] E. Strelcenia and S. Prakoonwit, “A New GAN-based Data Augmentation Method for Handling Class Imbalance in Credit Card Fraud Detection,” 2023, doi: 10.1109/spin57001.2023.10116543.

[24] J. Feng, “RBPR: A hybrid model for the new user cold start problem in recommender systems,” *Knowl.-Based Syst.*, vol. 214, no. Query date: 2024-06-11 15:47:26, 2021, doi: 10.1016/j.knosys.2020.106732.

[25] A. D. Riyanto, A. M. Wahid, and A. A. Pratiwi, “ANALYSIS OF FACTORS DETERMINING STUDENT SATISFACTION USING DECISION TREE, RANDOM FOREST, SVM, AND NEURAL NETWORKS: A COMPARATIVE STUDY,” *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, Art. no. 4, Jul. 2024, doi: 10.52436/1.jutif.2024.5.4.2188.