Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 2 April 2025, pp: 598-608
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Simulation of Student Study Group Formation Design Using K-Means Clustering

Yudistira Ardi Nugraha Setyawan Putra[1*], Hendro Margono[2]

[1]Department of Human Resource Development, Faculty of Graduate School, Airlangga University, Indonesia
[2]Department of Information and Library Science, Faculty of Social and Political Sciences, Airlangga University, Indonesia

E-Mail: [1]yudistira.ardi.nugraha-2023@pasca.unair.ac.id, [2]hendro.margono@fisip.unair.ac.id

Received Nov 29th 2024; Revised Feb 16th 2025; Accepted Feb 22th 2025; Available Online Mar 21th 2025, Published Jan 21th 2025
Corresponding Author: Yudistira Ardi Nugraha Setyawan Putra

Abstract

This research focuses on developing a simulation model for forming student study groups using an enhanced K-Means algorithm, addressing the challenge of optimizing group dynamics to improve learning outcomes. By analyzing the effectiveness of the formed study groups through Root Mean Square Error (RMSE) after dimensionality reduction with various regression models—including Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor—we aim to provide educators with a robust tool for assessing group configurations. The study identifies four distinct clusters, revealing that "Previous_Score" and "Attendance" are critical variables, achieving a highest Silhouette Score of 0.64 with five selected features. The ridge regression model also yielded a low RMSE of 0.045, explaining 72.39% of the variance in "Exam_Score." The findings suggest that targeted interventions tailored to each cluster—yellow, purple, blue, and green—can enhance academic outcomes by addressing specific student needs. This data-driven approach optimizes group dynamics and fosters a more inclusive learning environment, enhancing academic performance and cultivating essential social skills. The study underscores the potential of machine learning techniques in education and suggests avenues for future research into alternative clustering methods and their long-term impact on student engagement and success.

Keyword: Academic Performance, K-Means Clustering, Machine Learning, Ridge Regression, Student Group Formation

1. INTRODUCTION

Effective learning relies heavily on forming appropriate study groups that foster collaboration and enhance understanding of the material, allowing students to share diverse perspectives and resources. Forming appropriate study groups significantly impacts student engagement and motivation by fostering a sense of belonging, enhancing social skills, and improving academic performance. Research indicates that study groups can create a supportive community, which is crucial for enhancing motivation and accountability, particularly in online learning environments. For instance, a study found that students participating in study-together groups reported a higher sense of belonging and improved academic performance, especially among those with lower academic preparation and motivation [1]. Teacher learning groups (TLGs) also highlight the importance of social learning, where factors such as autonomous choices, sharing, and support are key to maintaining motivation [2]. The size of the study group, whether small or large, does not inherently affect motivation; rather, the method of delivery and the student's responsibility play more significant roles [3]. Inclusive study group formation, which allows for continuous refinement and matching based on student needs, has been shown to provide positive experiences and higher exam scores, particularly benefiting students from underrepresented backgrounds [4]. Additionally, grouping students with similar engagement levels can enhance learning achievement and satisfaction, particularly for low and high-engagement students, respectively [5].

Conventional methods in forming study groups often do not optimally consider various student characteristics. Conventional methods of forming study groups often face several weaknesses that hinder effective learning outcomes. One primary issue is the lack of consideration for the compatibility and aptness of group members, which is crucial for achieving productive collaborative learning. Traditional approaches may not adequately account for the diverse static and dynamic characteristics of students, leading to suboptimal

group dynamics and learning experiences [6]. Additionally, the selection of inappropriate tools for group formation can result in poorly structured groups, which may lead to student dissatisfaction and reduced learning effectiveness. This is exacerbated by the potential for bullying tendencies if group compositions do not consider social dynamics and student preferences [7]. Furthermore, conventional methods often fail to establish clear learning outcomes and structured tasks, which are essential for fostering critical thinking and problem-solving skills within study groups. Without these elements, students may not engage effectively, and the group may not function optimally [8]. Another significant drawback is the tendency for small groups to not continue collaboration outside of class, which limits the potential for sustained learning and development. Additionally, grading based on group performance can create perceptions of unfairness, particularly if there is an imbalance in contributions among group members [9]. Finally, traditional group formation methods often overlook the need for systematic reformation based on individual skills and personalities, which can significantly impact learning outcomes. Empirical studies suggest that reorganizing groups during a course can lead to improved educational results, highlighting the importance of dynamic group management [10]. Overall, these weaknesses underscore the need for more sophisticated and adaptive group formation strategies to enhance collaborative learning outcomes. Innovative approaches that prioritize individual learning styles and preferences can lead to more effective collaboration, fostering an environment where all students feel valued and empowered to contribute.

With the increasing complexity of available student data and advances in machine learning, there is an opportunity to use the K-Means algorithm to group students based on relevant data [11]. The application of the K-Means algorithm for student grouping has been explored across various educational contexts, demonstrating its versatility and effectiveness. In the context of higher education, K-Means has been used to cluster students based on their areas of expertise, such as Software Engineering and IT Entrepreneurship, by analyzing their course scores. This approach successfully grouped students into distinct clusters, aiding in the identification of their strengths and potential thesis topics [12]. Similarly, in elementary education, K-Means has been employed to group students receiving financial aid from the Smart Indonesia Program (PIP) based on socio-economic factors, ensuring a fair distribution of resources [13]. In military education, K-Means was used to categorize students at the Pusdikjas institution into clusters based on performance metrics, which helped in assessing student capabilities and tailoring educational strategies [14]. However, traditional K-Means clustering can sometimes result in imbalanced group sizes. Generally, limitations in such studies may include challenges in selecting the optimal number of clusters, sensitivity to outliers, and assumptions about data distribution.

This research focuses on developing a simulation model for forming student study groups using an enhanced K-Means algorithm, while also analyzing the effectiveness of the formed study groups through RMSE (Root Mean Square Error) after dimensionality reduction using models such as such as Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor. The simulation model aims to provide educators with a robust tool for assessing group dynamics, enabling them to make more informed decisions about student placements based on individual learning styles and performance metrics. By leveraging this model, educators can identify the most effective group configurations that foster collaboration and maximize each student's strengths, thereby creating a more inclusive and supportive learning environment. This approach not only enhances academic performance but also cultivates essential social skills and teamwork among students, preparing them for future collaborative endeavors in both educational and professional settings.

2. RESEARCH METHODS

This research adopts a structured methodological approach as shown in Figure.1, using a dataset titled "Student Performance Factor", available on the Kaggle platform [15]. This dataset contains relevant academic information about students for analysis. The next stage involves data preprocessing, during which the data is cleaned and prepared to ensure optimal quality and consistency. Once the data is ready, the data modeling process is conducted by applying the K-Means algorithm through a series of steps designed to enhance its performance. We performed dimensionality reduction using models such as Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor to eliminate irrelevant columns based on RMSE. The formula for RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\acute{y}_i - y_i)^2}{n}} \quad (1)$$

Where, $\acute{y}_i$ are predicted value, y_i are observed value and n is thenumber observations.

Dimensionality reduction techniques using various regression models, including Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost, have been explored in literature. These models are evaluated based on their performance metrics, particularly RMSE, to assess their effectiveness in reducing irrelevant features. The goal of using RMSE in the context of dimensionality reduction is to identify which features contribute the least to the model's predictive

power, allowing for their removal [16]. In the realm of data analysis, dimensionality reduction has become a critical technique for enhancing model performance, particularly when dealing with high-dimensional datasets. Various regression models, including Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor, have been employed to eliminate irrelevant features based on performance metrics such as Root Mean Square Error (RMSE). For instance, studies have demonstrated that Lasso Regression, with its inherent feature selection capability, effectively reduces dimensionality by shrinking less important feature coefficients to zero, thereby simplifying the model without sacrificing predictive accuracy [17]. Similarly, Ridge Regression has been shown to mitigate multicollinearity issues in high-dimensional data, allowing for more stable estimates and improved model interpretability [18].

Moreover, ensemble methods like Random Forest and Gradient Boosting have gained traction for their ability to handle irrelevant features through their built-in feature importance measures. Research indicates that these models can significantly enhance predictive performance by focusing on the most relevant features while discarding those that contribute little to the outcome [19]. XGBoost, in particular, has been recognized for its efficiency and effectiveness in large datasets, often outperforming traditional regression techniques by leveraging advanced regularization techniques to prevent overfitting and improve generalization [20]. However, the process of dimensionality reduction is not without its challenges. Selecting the optimal number of features can be complex, and reliance on RMSE as a sole metric may overlook other important aspects of model performance, such as interpretability and computational efficiency. Additionally, the sensitivity of certain models to outliers can skew results, leading to suboptimal feature selection [21]. Overall, while dimensionality reduction using these regression models offers significant advantages in improving model performance, careful consideration must be given to the selection process and the inherent limitations of each approach.

The final step is evaluation using the Silhouette Score, a metric that measures the quality of the clusters generated by K-Means by assessing how well the data within each cluster is grouped and how distinctly separated the clusters are. This approach aims to produce clusters that are both relevant and meaningful in understanding the factors influencing student performance.

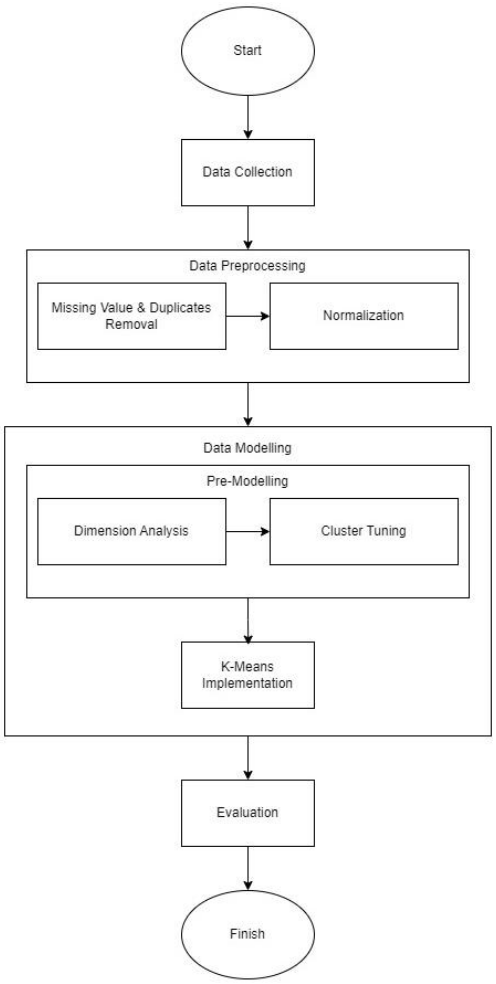

Figure 1. Research Methodology

2.1. Data Collection

In this research, the data used consists of secondary data obtained from the Kaggle website under the title "Student Performance Factors". This dataset contains 6,607 raw data entries across 20 variable columns. Each column represents different factors that can influence student performance, such as socioeconomic background, study habits, and interactions within the educational environment, as shown in Table 1.

Table 1. Columns Description

Column	Description
Hours_Studied	Number of hours spent studying per week (int64)
Attendance	Percentage of classes attended (int64)
Parental_Involvement	Level of parental involvement in the student's education (object)
Access_to_Resources	Availability of educational resources (object)
Extracurricular_Activities	Participation in extracurricular activities (object)
Sleep_Hours	Average number of hours of sleep per night (int64)
Previous_Scores	Scores from previous exams (int64)
Motivation_Level	Student's level of motivation (int64)
Internet Access	Availability of internet access (int64)
Tutoring_Sessions	Number of tutoring sessions attached per month (int64)
Family_Income	Family income level (object)
Teacher_Quality	Quality of teacher (object)
School_Type	Type of school attended (object)
Peer_Influence	Influence of peers on academic performance (object)
Physical_Activity	Average number of hours of physical activity per week (int64)
Learning_Disabilities	Presence of learning disabilities (object)
Parental_Education_Level	Highest education level of parents (object)
Distance_from_Home	Distance from home to school (object)
Gender	Gender of the student (object)
Exam_Score	Final exam score (int64)

Utilizing secondary data enables researchers to analyze and understand the patterns and relationships among these factors without the need for direct data collection, thereby accelerating the research process and offering deeper insights into the determinants of student academic performance. For example, the histogram in Figure 2, represents the distribution of study hours from a dataset, visually indicating how frequently each range of study hours occurs among participants. In the X-axis (Hours Studied), it ranges from 0 to 40 hours, it represents the total number of hours studied by the individuals in the study. For Y-Axis (Frequency), it reflects the count of participants who fall within each range of study hours. The values increase up to a maximum near 900, indicating a significant number of individuals studied within a specific range. The histogram exhibits a roughly normal distribution shape with a slight positive skew. The peak is observed around 15-20 hours, suggesting that most participants studied within this range.

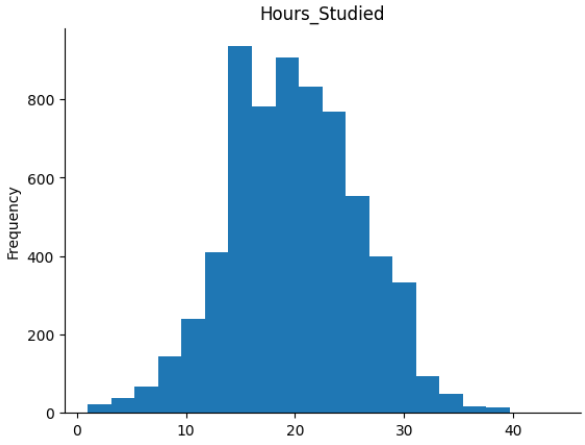

Figure 2. Data Distribution for Student Hours Studied

2.2. Data Preprocessing

At this stage, an in-depth analysis is conducted to ensure the quality and integrity of the data used in the analytical model. This process begins with identifying anomalies, such as missing values and duplicate data. Missing values can compromise the model's accuracy and must be handled carefully, either by imputing appropriate estimates or removing incomplete entries [22]. Similarly, duplicate data can introduce bias into the

analysis, making it essential to detect and remove them to ensure the dataset reflects unique and relevant information, see Table 2 [23].

Table 2. Identified Missing Value and Duplicate

Column	Missing Value	Duplicate
Hours_Studied	0	0
Attendance	0	0
Parental_Involvement	0	0
Access_to_Resources	0	0
Extracurricular_Activities	0	0
Sleep_Hours	0	0
Previous_Scores	0	0
Motivation_Level	0	0
Internet Access	0	0
Tutoring_Sessions	0	0
Family_Income	0	0
Teacher_Quality	78	0
School_Type	0	0
Peer_Influence	0	0
Physical_Activity	0	0
Learning_Disabilities	0	0
Parental_Education_Level	90	0
Distance_from_Home	67	0
Gender	0	0
Exam_Score	0	0
Total	235 (3.5%)	0(0%)

Out of the 6,607 entries, 235 were marked as having missing values. Since this accounts for only 3.5% of the total data, these entries were subsequently removed [24]. After addressing these anomalies, the process continues with normalizing the values for each variable to ensure that differences in data scales do not affect the analysis results and enhance clustering performances. Categorical variables are transformed using one-hot encoding, which converts categories into a binary format, making them suitable for machine learning models [25]. Numeric variables are normalized using the MinMaxScaler, which scales values to a specified range, typically between 0 and 1 [26].

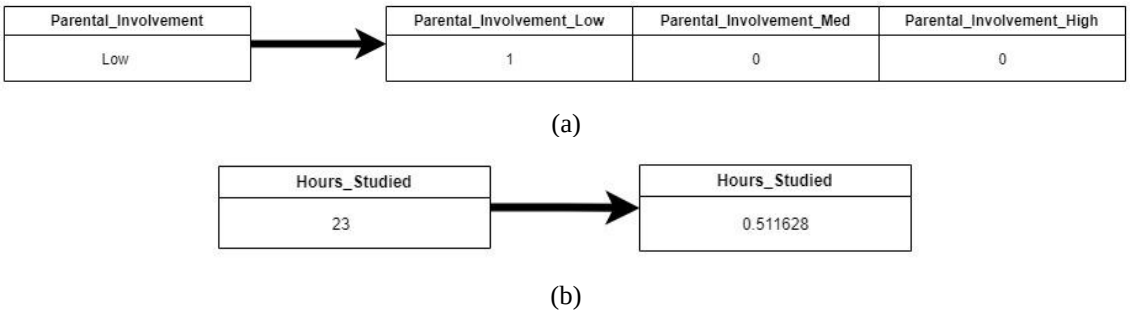

Figure 3. Illustrations of (a) One-Hot Encoding and (b) Min Max Scaler

This preprocessing step is crucial for improving the accuracy and efficiency of machine learning algorithms, allowing them to learn patterns more effectively from the data. This careful preparation of data not only facilitates better model performance but also ensures that the insights drawn from the analysis are both meaningful and actionable in real-world applications.

2.3. Data Modelling

At the Data Modeling stage, dimensional analysis is carried out to assess whether reducing data dimensions is needed to increase the efficiency and effectiveness of the model to be built [27]. Previously, the normalization process for categorical variables resulted in additional data dimensions, namely in the form of widening the number of columns from 20 columns to 41 columns as illustrated in Fig.3. This research integrates several methods to determine the best steps in handling large-dimensional data for clustering needs. This process involves evaluating several machine learning models that have proven their performance in processing numerical data, such as Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor, with assigning “Exam_Score” as the target variable. The results of this analysis help in identifying the most suitable model for predicting exam scores,

allowing to make informed decisions about feature selection and dimensionality reduction techniques that could enhance predictive accuracy. Each of these models offers unique advantages and can be selected based on the specific characteristics of the dataset, including its size, complexity, and the nature of the relationships within the data. Choosing the appropriate model requires careful consideration of factors such as interpretability, computational efficiency, and the potential for overfitting, which can significantly impact the overall performance of the predictive analytics [28]. The approach also includes various methods, namely without feature selection (FS), Filter Method, Wrapper Method, and Embedded Method, to determine the relevance and contribution of each variable in the model. Evaluating these methods in conjunction with the chosen regression models can lead to more effective feature selection strategies, ultimately enhancing model accuracy and robustness while minimizing unnecessary complexity [29]. After going through the evaluation process, the next step is to choose the most appropriate cluster number based on elbow method, which is not only able to provide accurate results but can also identify the variables that most influence student academic performance [30]. Thus, this stage focuses not only on the accuracy of predictions, but also on a deeper understanding of the factors that influence academic achievement, which can serve as a basis for decision making and future improvements.

2.4. Evaluation

In the Evaluation stage, the analysis process focuses on measuring the quality of the resulting clusters through the use of silhouette scores. Silhouette score is an effective metric for assessing how well each data point is classified in its cluster compared to other clusters [31]. The silhouette score value ranges from -1 to 1, where a value close to 1 indicates that the data point is in the right cluster and is well separated from other clusters. Conversely, values close to -1 indicate that the data points may have been grouped into the wrong cluster. By using the silhouette score, we can evaluate and compare various cluster configurations, and determine the optimal number of clusters. This evaluation process is critical to ensure that the clustering model built not only produces well-separated groups, but also provides meaningful insights from the analyzed data, thereby supporting better decision making in the future.

3. RESULTS AND DISCUSSION

3.1. Results

The model results indicate that using four clusters and five selected features yields the highest Silhouette Score, with a value of 0.64. Among the clusters formed, the cluster graph of the variable combination "Previous_Score" and "Attendance" exhibits the most distinct pattern compared to other variable combinations. This suggests that grouping students based on "Previous_Score" and "Attendance" produces the most optimal student clusters.

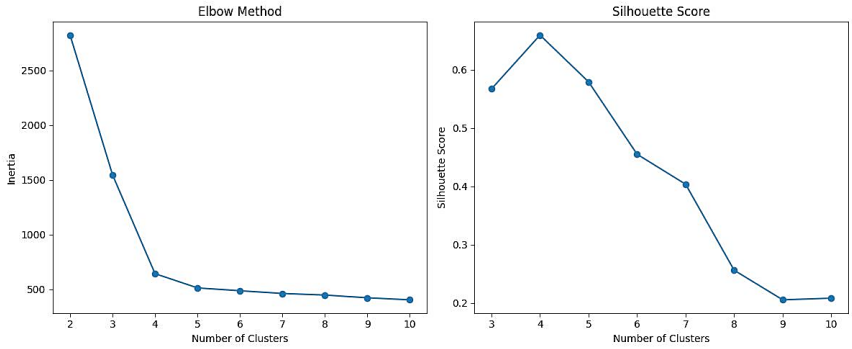

Figure 4. Results of Elbow Method and Silhouette Score

This finding underscores the importance of these two variables in understanding student performance and highlights their potential to inform targeted interventions aimed at improving academic outcomes. The five selected features were chosen based on the top five most significant coefficient values. Additionally, the results of the RMSE performance comparison during the dimensional analysis stage indicate that using the ridge regression model without feature selection produced the smallest RMSE value of 0.045, making it the most effective approach for identifying the relationship between variables and the target variable.

Table 3. Top 5 Dimension Analysis RMSE Comparisons for the Reduction Approach

Method	RMSE
Ridge Regression (No FS)	0.044533
Linear Regression (No FS)	0.044550
Gradient Boosting Regressor (No FS)	0.047281

Method	RMSE
Ridge Regression (Filter FS)	0.051047
Linear Regression (Filter FS)	0.051047

The smallest value of RMSE indicates that the combination of ridge regression without feature selection offers the best model for investigating the relationship between the variables and the target variable, "Exam_Score" [32]. In addition, the coefficient of determination value from ridge regression without feature selection is 0.7239 for Ridge Regression, indicating that the model can explain around 72.39% of the data variation in the target variable ("Exam_Score") based on the independent variables used. A coefficient of determination value close to one indicates that the model has quite good predictive ability [33], but around 27.61% of the variation cannot be explained by the model. This could be caused by other variables not included in the model, noise, or a mismatch between the model and the data [34]. In this context, the model successfully explains most of the relationship patterns between the independent and target variables. The other results of this analysis yield coefficient values that describe the degree of influence each variable has on the target variable. These coefficients are represented on a scale from -1 (indicating a significant negative effect) to 1 (indicating a significant positive effect). Based on the analysis, five variables were identified as influencing "Exam_Score". These variables are "Attendance", "Hours_Studied", "Previous_Score", "Tutoring_Sessions" and "Access_to_Resources_Low".

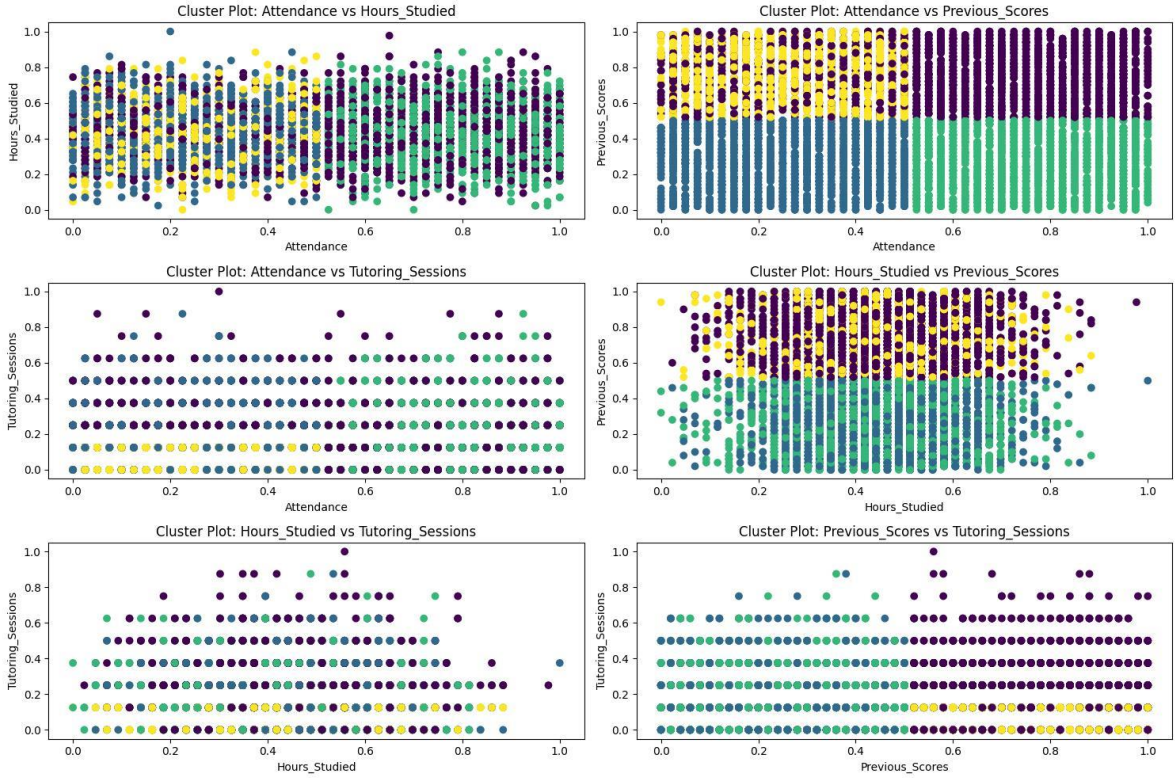

Figure 5. Cluster Plot Comparisons

Based on the cluster plot of the "Previous Score" and "Attendance" variables in Fig.5, four distinct groups of students were identified, which can be utilized to tailor strategies for improving their academic performance. First, the yellow group consists of students with low to moderate attendance but high previous test scores. Even if they have demonstrated good academic ability in the past, low attendance rates can be an indicator of potential decline in performance in the future. Students in this group may need additional motivation or support to improve their attendance, so as to maintain or improve their performance in future exams. Identifying the underlying reasons for their absenteeism is crucial, as addressing these factors can lead to tailored interventions that foster both attendance and academic success. By implementing strategies such as personalized mentoring, flexible scheduling, and engaging learning activities, educators can create an environment that encourages regular attendance and reinforces students' academic strengths.

Second, the purple group consists of students with low to high attendance rates, but consistently high previous test scores. Students in this group demonstrate strong academic potential, and varying attendance rates may indicate that they have a flexible approach to learning. However, to ensure their success, it is important to maintain a good attendance rate, as high attendance often contributes to better understanding of the material

and better exam results. To support these students, schools should consider providing additional resources and incentives that motivate regular attendance while recognizing their academic capabilities. One effective strategy could be the implementation of mentorship programs, where students can receive guidance and encouragement from peers or teachers to foster a sense of belonging and accountability within the school community.

Third, the blue group consists of students with low to moderate attendance rates and previous test scores that range from low to moderate. This group may face challenges both in terms of motivation and understanding the material. Students in this group may need additional support, such as academic tutoring or remedial programs, to help improve their understanding of course material and, in turn, improve their test scores. Providing personalized learning plans tailored to each student's unique needs can further enhance their academic experience, ensuring that they receive the specific resources and attention necessary for their individual growth.

Last, the green group consists of students with high attendance but low previous test scores. Although students in this group demonstrate a good commitment to attendance, low test scores indicate that they may have difficulty understanding the material or applying the knowledge they have learned. Interventions that focus on strengthening understanding of exam concepts and techniques can be especially beneficial for students in this group. By incorporating targeted tutoring sessions and practice exams, educators can help these students build confidence in their abilities while also improving their performance on assessments. This tailored approach not only addresses the academic challenges faced by these students but also fosters a supportive learning environment that encourages persistence and resilience.

Both the yellow group and purple group are not well separated, even a silhouette score of 0.64 indicates a relatively good clustering result. However, there may still be some overlap or ambiguity at the cluster boundaries, which could be further refined. In particular, exploring different clustering algorithms or fine-tuning the parameters of the current model may enhance the clarity and separation of the clusters, leading to even more accurate representations of the underlying data structure [35] and testing various distance metrics or incorporating additional features could also provide new insights, potentially revealing hidden patterns that were not apparent with the current approach [36].

3.2. Discussion

The analysis yielded valuable insights into student clustering and performance prediction using key academic features. The optimal model constituted four clusters formed based on five selected features, achieving the highest Silhouette Score of 0.64. Among the analyzed features, "Previous_Score" and "Attendance" emerged as critical determinants of student performance, underscoring their potential utility in tailoring educational interventions. The selected features, heavily weighted by their significance coefficients, were “Attendance”, “Previous_Score”, “Hours_Studied”, “Tutoring_Sessions”, “Access_to_Resource_Low”. The clustering analysis illustrated distinct patterns, particularly highlighting that students grouped based on "Previous_Score" and "Attendance" exhibit substantial differences in performance potential. This finding aligns with existing literature [37] which emphasizes the role of attendance in academic success and suggests that prior performance often serves as a predictive indicator. The identification of these variables as significant contributors to student performance can inform educational strategies aimed at improving outcomes.

The four clusters identified in this study present unique opportunities for tailored interventions. Research indicates that tailored interventions can significantly enhance student performance by addressing attendance patterns and previous academic scores. Strategies such as personalized mentoring, flexible scheduling, and targeted tutoring are effective in engaging students and improving their academic outcomes [38]. Understanding individual needs is crucial for successful intervention. Studies have shown that personalized mentoring programs can lead to improved student engagement and academic performance. For instance, a meta-analysis found that mentoring relationships positively impact students' motivation and self-efficacy, which are critical for maintaining attendance and academic success [39]. Secondly, Research supports the idea that flexible scheduling can accommodate students' diverse needs, particularly those facing external challenges [40]. A study highlighted that students with access to flexible learning environments reported higher satisfaction and engagement, leading to improved attendance rates. Last, evidence suggests that targeted tutoring programs can effectively address comprehension issues among students. A randomized controlled trial demonstrated that students receiving tailored tutoring showed significant improvements in their academic performance compared to those who did not receive such support [41].

The integration of these research findings underscores the importance of implementing tailored interventions that consider the unique needs of different student groups. By leveraging evidence-based strategies, educators can create supportive environments that enhance both attendance and academic performance, ultimately contributing to student success.

4. CONCLUSION

This research demonstrates the effectiveness of K-Means clustering analysis in understanding the factors influencing student performance. The results indicate that utilizing four clusters and five selected features yields the highest Silhouette Score of 0.64, with "Previous_Score" and "Attendance" emerging as the most significant variables in forming distinct and meaningful student groups. These clusters provide valuable insights for tailoring targeted interventions aimed at enhancing academic outcomes.

The ridge regression model, which was applied without feature selection, achieved the lowest RMSE value of 0.045 and a coefficient of determination of 0.7239. This indicates that the model explains 72.39% of the variance in "Exam_Score," while the remaining 27.61% may be attributed to unmodeled factors or noise. The identified significant features—"Attendance," "Hours_Studied," "Previous_Score," "Tutoring_Sessions," and "Access_to_Resources_Low"—highlight key areas for educational focus and support.

The analysis of student clusters further revealed actionable insights, with each group—yellow, purple, blue, and green—requiring specific interventions ranging from enhancing attendance to providing targeted academic support. While the clustering results are relatively robust, the overlap between the yellow and purple groups suggests opportunities for refinement.

However, the promising nature of these clustering results is tempered by the need for further refinement. The ambiguity at the boundaries of these clusters indicates potential for improvement. Future research could explore alternative clustering algorithms or fine-tune the parameters of the current model to enhance the clarity and separation of the clusters. Additionally, testing various distance metrics or incorporating additional features may reveal hidden patterns that were not apparent in this analysis.

Moreover, the coefficient of determination value of 0.7239 suggests that while the model explains a significant portion of the variation in "Exam_Score," there remains a substantial 27.61% of the variation that is unexplained. This gap points to the presence of other influential variables not included in the model, as well as potential noise or mismatches between the model and the data. Future studies should consider incorporating a broader range of variables, such as socio-economic factors, learning styles, and psychological aspects, to provide a more comprehensive understanding of the factors influencing academic performance.

In summary, this study underscores the potential of data-driven approaches in educational settings to inform policies and strategies that address diverse student needs, ultimately fostering improved academic performance. The findings suggest that educational institutions can leverage these insights to develop targeted interventions that enhance student engagement and success. Further research in this area could lead to more refined models and a deeper understanding of the complex factors that influence student achievement.

REFERENCES

[1] X. Zhou, Q. Li, D. Xu, A. Holton, and B. Sato, “The promise of using study-together groups to promote engagement and performance in online courses: Experimental evidence on academic and non-cognitive outcomes,” Internet and Higher Education, Sep. 2023, doi: 10.1016/j.iheduc.2023.100922.

[2] E. Vrieling-Teunter, N. de Vries, P. Sins, and M. Vermeulen, “Student motivation in teacher learning groups,” European Journal of Teacher Education, Jun. 2022, doi: 10.1080/02619768.2022.2086119.

[3] N. Davidovitch and R. Yavich, “Study group size, motivation and engagement in the digital era,” Problems of education in the 21st century, Jun. 2023, doi: 10.33225/pec/23.81.361.

[4] “Inclusive Study Group Formation At Scale,” Feb. 2022, doi: 10.48550/arxiv.2202.07439.

[5] V. Abou-Khalil and H. Ogata, “Homogeneous Student Engagement: A Strategy for Group Formation During Online Learning,” Aug. 2021, doi: 10.1007/978-3-030-85071-5_6.

[6] N. Sarode and J. W. Bakal, “Toward Effectual Group Formation Method for Collaborative Learning Environment,” Jan. 2021, doi: 10.1007/978-981-15-8677-4_29.

[7] K. Lee, J. Ko, C. Jwa, and J. Cho, “Development of Grouping Tool for Effective Collaborative Learning,” Journal of Digital Convergence, Jan. 2018, doi: 10.14400/JDC.2018.16.7.243.

[8] W. D. Linn, K. C. Lord, C. Y. Whong, and E. G. Phillips, “Developing effective study groups in the quest for the ‘Holy Grail’: critical thinking.,” The American Journal of Pharmaceutical Education, Oct. 2013, doi: 10.5688/AJPE778180.

[9] V. H.-I. Chi and P. Kadandale, “All Groups Are Not Created Equal: Class-Based Learning Communities Enhance Exam Performance and Reduce Gaps,” CBE- Life Sciences Education, Sep. 2022, doi: 10.1187/cbe.21-09-0240.

[10] A. Mujkanovic and A. Bollin, “Improving learning outcomes through systematic group reformation: the role of skills and personality in software engineering education,” International Conference on Software Engineering, May 2016, doi: 10.1145/2897586.2897615.

[11] P. I. Ciptayani, K. C. Dewi, and I. W. B. Sentana, “Student grouping using adaptive genetic algorithm,” in International Electronics Symposium, Sep. 2016. doi: 10.1109/ELECSYM.2016.7861034.

[12] Y. Y. L. Yuyun, C. R. T. Sinaga, M. Nugroho, and M. Ridha, “K-Means Algorithm for Clustering Students Based on Areas of Expertise (A Case Study),” Jun. 2024, doi: 10.62123/aqila.v1i1.23.

[13] J.-P. Huang, P.-C. Wang, and R. M. F. Lubis, “The Process of Grouping Elementary School Students

Receiving PIP Assistance uses the K-Means Algorithm," Bulletin of Informatics and Data Science, Nov. 2023, doi: 10.61944/bids.v2i2.78.
[14] I. W. Pramudjianto, A. K. Ningsih, and A. Komarudin, "Grouping Education Students at Pusdikjas Institutions of The TNI-AD's Disjasad Using the K-Means Clustering Method," Oct. 2023, doi: 10.55324/enrichment.v1i7.64.
[15] P. Data, "Student Performance Factors," Kaggle.com, 2024.https://www.kaggle.com/datasets/lainguyn123/student-performance-factors
[16] A. Laakel Hemdanou, M. Lamarti Sefian, Y. Achtoun, and I. Tahiri, "Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models," Computers and Education: Artificial Intelligence, vol. 7, p. 100301, Dec. 2024, doi: https://doi.org/10.1016/j.caeai.2024.100301.
[17] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996, Available: https://www.jstor.org/stable/2346178
[18] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, vol. 42, no. 1, p. 80, Feb. 2000, doi: https://doi.org/10.2307/1271436.
[19] L. Breiman, "Random Forests," Jan. 2001. Available: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf
[20] T. Chen and C. Guestrin, "XGBoost: a Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, pp. 785–794, 2016, doi: https://doi.org/10.1145/2939672.2939785.
[21] V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence Review, vol. 22, no. 2, pp. 85–126, Oct. 2004, doi: https://doi.org/10.1007/s10462-004-4304-y.
[22] F. Husson and J. Josse, "Handling missing values in multiple factor analysis," Food Quality and Preference, Dec. 2013, doi: 10.1016/J.FOODQUAL.2013.04.013.
[23] S. Sharma, Y. Zhang, J. Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney, "Data Augmentation for Discrimination Prevention and Bias Disambiguation," in National Conference on Artificial Intelligence, Feb. 2020. doi: 10.1145/3375627.3375865.
[24] D. Panda, "Does data cleaning disproportionately affect autistics," Autism, Feb. 2018, doi: 10.1177/1362361316673566.
[25] P. Cerda and G. Varoquaux, "Encoding high-cardinality string categorical variables," arXiv: Learning, Jul. 2019, doi: 10.1109/TKDE.2020.2992529.
[26] B. Wang et al., "A Normalized Numerical Scaling Method for the Unbalanced Multi-Granular Linguistic Sets," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Apr. 2015, doi: 10.1142/S0218488515500099.
[27] A. Matuszak, "Dimensional Analysis can Improve Equations of the Model," Procedia Engineering, Jan. 2015, doi: 10.1016/J.PROENG.2015.06.174.
[28] G. C. Cawley, "Over-fitting in model selection and its avoidance," 2012. doi: 10.1007/978-3-642-34156-4_1.
[29] Y. Sun, J. Yao, and S. Goodison, "Feature selection for nonlinear regression and its application to cancer research," in SIAM International Conference on Data Mining, Jan. 2015. doi: 10.1137/1.9781611974010.9.
[30] M. Musso, E. Kyndt, E. Cascallar, and F. Dochy, "Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks," Aug. 2013, doi: 10.14786/FLR.V1I1.13.
[31] L. Lovmar, A. Ahlford, M. Jonsson, and A.-C. Syvänen, "Silhouette scores for assessment of SNP genotype clusters," BMC Genomics, Mar. 2005, doi: 10.1186/1471-2164-6-35.
[32] S. Paul and P. Drineas, "Feature selection for ridge regression with provable guarantees," Neural Computation, Apr. 2016, doi: 10.1162/NECO_A_00816.
[33] W. S. Dong, C. H. Tian, Y. Wang, J. Yan, and C. Zhang, "Method and apparatus for evaluating predictive model," Jun. 25, 2014
[34] J. Peters, D. Janzing, and B. Schölkopf, "Identifying Cause and Effect on Discrete Data using Additive Noise Models," in International Conference on Artificial Intelligence and Statistics, Mar. 2010.
[35] S. Huang, F. Wei, L. Cui, X. Zhang, and M. Zhou, "Unsupervised Fine-tuning for Text Clustering," in International Conference on Computational Linguistics, Dec. 2020. doi: 10.18653/V1/2020.COLING-MAIN.482.
[36] T. Yoshida, I. Takeuchi, and M. Karasuyama, "Learning Interpretable Metric between Graphs: Convex Formulation and Computation with Graph Mining," in Knowledge Discovery and Data Mining, Jul. 2019. doi: 10.1145/3292500.3330845.
[37] K. Al Hazaa et al., "The effects of attendance and high school GPA on student performance in first-year undergraduate courses," Cogent Education, vol. 8, no. 1, p. 1956857, Jan. 2021, doi:

https://doi.org/10.1080/2331186x.2021.1956857.
[38] S. White, L. Groom-Thomas, and S. Loeb, “Undertaking complex but effective instructional supports for students: A systematic review of research on high-impact tutoring planning and implementation,” doi: https://doi.org/10.26300/wztf-wj14.
[39] D. L. DuBois, B. E. Holloway, J. C. Valentine, and H. Cooper, “Effectiveness of Mentoring Programs for Youth: A Meta-Analytic Review,” American Journal of Community Psychology, vol. 30, no. 2, pp. 157–197, Apr. 2002, doi: https://doi.org/10.1023/a:1014628810714.
[40] M. Colasante, J. Bevacqua, and S. Muir, “Flexible hybrid format in university curricula to offer students in-subject choice of study mode: An educational design research project,” Journal of University Teaching and Learning Practice, vol. 17, no. 3, pp. 119–136, Jul. 2020, doi: https://doi.org/10.53761/1.17.3.9.
[41] K. E. Cortes, K. Kortecamp, S. Loeb, and C. D. Robinson, “A scalable approach to high-impact tutoring for young readers,” Learning and Instruction, vol. 95, pp. 102021–102021, Sep. 2024, doi: https://doi.org/10.1016/j.learninstruc.2024.102021.