Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

Data-Driven Student Group Formation for Group Investigation: A K-Medoids Clustering Approach in Cooperative Learning

Salma Alyasyifa*[1], Oktariani Nurul Pratiwi[2], Irfan Darmawan[3]

[1,2,3]Information System, Telkom University, Indonesia

Email: [1]almalyas@student.telkomuniversity.ac.id

Received : May 22, 2025; Revised : Jul 19, 2025; Accepted : Jul 30, 2025; Published : Oct 23, 2025

Abstract

Group Investigation (GI) is a widely used cooperative learning strategy in higher education, but challenges such as large class sizes and diverse student profiles complicate manual group formation. Previous studies have applied clustering algorithms like K-Means, yet K-Medoids, which is robust to noise, remain underexplored for group formation, especially GI. This study proposed a data-driven approach using the K-Medoids clustering algorithm to create student groups that are both interest-aligned and heterogeneous in profile, which enhancing the effectiveness of GI activities. Employing the Knowledge Discovery in Databases (KDD) framework, the process included data selection, preprocessing, transformation, three grouping processes, and evaluation were performed. In grouping process students were initially grouped by interest, clustered using K-Medoids with various distance measures tested, and finally, groups were adjusted to balance homogeneity and diversity. In grouping stage 2, clustering with Euclidean distance and PCA achieved the highest Silhouette Score, indicating superior grouping quality. The result of heterogeneity group of students evaluated with Gower dissimilarity shows that the method produces internally diverse yet cohesive interest groups, supporting GI goals.

Keywords : *Cooperative Learning, Group Investigation, K-Medoids, Student Group Formation.*

This work is an open access article and licensed under a Creative Commons Attribution-Non Commercial 4.0 International License

1. INTRODUCTION

Group Investigation (GI), a model of cooperative learning, is an active learning approach that involves grouping students to collaboratively achieve shared goals. This method helps students construct their own knowledge, develop transferable skills, and enhance both engagement and academic outcomes [1], [2]. GI is widely used in higher education settings [3], [4], [5]. Unlike traditional methods that focus on solving a specific problem or task, the Group Investigation approach encourages students to work together to explore and research a particular topic in depth [4].

Group formation is a critical component of effective collaborative learning [6]. Group formation methods are generally divided into two: self-selected teams and instructor-formed teams. Although self-selected teams often result in positive perceptions of communication and trust between members [7], [8] however, the effectiveness of self-selected teams depends on the nature of the task; they tend to perform better when the task requires minimal collaboration [8].

According to the comprehensive guide by Yael Sharan and Shlomo Sharan, Group Investigation model allows students to select topics based on their personal interests (self-selected). Afterward, the instructor forms heterogeneous groups based on specific characteristics such as academic performance, gender, and social personality traits (instructor-formed) [9], [10], [11]. Each group then conducts an investigation on their chosen topic and presents the results to the class.

However, the implementation of cooperative learning in higher education is not without challenges. Large class sizes, group formation, diverse student needs, manually balancing groups based

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

on various characteristics is not an easy task [12], [13], [14]. Therefore, algorithm-based approaches such as K-Medoids can be used to help form appropriate group compositions objectively and efficiently.

Automated group formation in cooperative learning commonly employs clustering and optimization techniques to enhance group diversity and improve learning outcomes. Several studies have explored clustering-based approaches: for instance, Kanika [15] applied the K-Modes algorithm to group students based on their conceptual understanding, while Nalli et al. [16] conducted a comparative analysis of clustering algorithms for Moodle plugin group formation and found that K-Means offered the best trade-off between accuracy and efficiency. Similarly, other researchers have utilized K-Means on behavioural data from Moodle to create heterogeneous student groups [17], [18].

Regression models have also been implemented such as research by Bousalem et.al [19] proposed a hybrid grouping strategy based on predicted student performance using regression models. While effective in balancing comfort and academic potential, the approach focuses mainly on achievement metrics. In parallel, optimization-based methods such as Particle Swarm Optimization [20] and Genetic Algorithms [21], [22] have been used to form groups by considering factors like personality traits and prior academic performance. Other methods include Magic Square Arrays [23] and Minimum Entropy Collaborative Groupings (MECG) [12], which uses complex network theory to improve group dynamics.

Despite these advancements, the use of the K-Medoids algorithm—which is more robust than K-Means in handling noise [24]—remains relatively underexplored in the context of student group formation. Additionally, limited research has integrated both student profile features and personal interests into the grouping process, especially within the framework of Group Investigation, a cooperative learning model that emphasizes group autonomy and diversity. This study addresses these gaps by proposing a K-Medoids-based approach that integrates both student interests and profile attributes to form heterogeneous groups, aligned with the pedagogical principles of the Group Investigation model.

2. METHOD

This research adopts a modified version of the Knowledge Discovery in Databases (KDD) framework. While data mining is often used interchangeably with KDD, technically, data mining is one stage within the overall KDD process, which involves discovering useful patterns or knowledge from dataset [25]. The KDD framework in this study guides the formation of student groups for the Group Investigation (GI) learning model. The methodology follows key stages: data selection, preprocessing, transformation, data mining, evaluation, and knowledge representation. This study follows a structured methodology as shown in Figure 1.

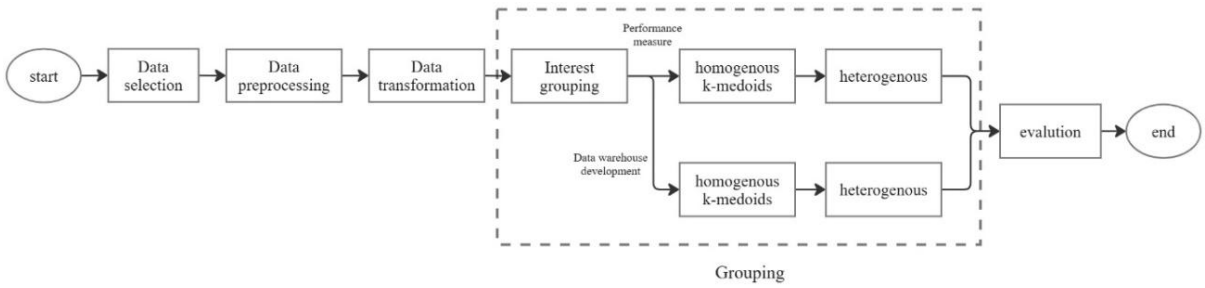

Figure 1. Research Methodology

The process begins with data selection based on student profile, data preprocessing, data transformation followed the requirement of K-Medoids data, grouping process with three stages: (1) students were first divided by their interest, (2) each subset was then clustered using the K-Medoids

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

algorithm, with the number of clusters selected based on the highest silhouette score, and (3) final groups were assembled by alternately selecting students from different clusters to ensure interest homogeneity and profile heterogeneity, and evaluation of grouping.

2.1. Data Selection

This study uses student profiles to guide data selection. A student profile is a structured representation that includes both explicit and implicit information about a student [26]. The data selection is based on the framework proposed by Hamim et al. [26], which identifies and organizes key global features of a student profile into ten aspects: personal identity, social identity, physical condition, academic, learning, cognitive, knowledge, psychological, and skills/interest. For this study, we adopted five aspects of the student profile—academic, personal identity, psychological, learning, and interest. These were selected based on their relevance and feasibility in a classroom setting. Other aspects, such as social identity and physical condition, were excluded due to data limitations. Group formation was applied in the DWBI course, where students selected one of two sub-topics (data warehouse development or performance measurement) to reflect their learning interests. Table 1. summarizes the variables and indicators used in this study based on the selected aspects of the student profile model by Hamim et al.

Table 1. Dataset description

Profile Aspect	Variable Name	Description	Scale/Value
Academic	Quiz1-3	All quizzes of DWBI course	0-10
	GPA	Cumulative GPA	0-4.00
	BASDAT	Final grade in the database system course, selected due to its relevance to DWBI	A-E
	DATA MINING	Final grade in the Data Mining course, selected due to its relevance to DWBI	A-E
Personal identity	Gender	Student's gender	Male/Female
	NIM	Unique student identifier	-
Psychological	MBTI	16 type MBTI classification	ISTJ, ISTP, etc.
Learning	VARK	Student's learning style	V, A, K
Interest	Minat	Student's topic choice in DWBI course	Data warehouse development / Performance measurement

2.2. Data Preprocessing and Transformation

Data preprocessing includes handling missing values, data integration, and data transformation or normalization. Among these, missing values are a common problem across many domains that deal with data. They can cause various issues, such as reduced model performance, biased outcomes, and difficulties in data analysis due to discrepancies between complete and incomplete records. A simple imputation approach involves replacing missing values with statistical estimates such as the mean, median, or mode of the observed values [27], [28]. The mean is typically used for numerical data, while the mode is commonly applied to categorical data [28].

Data transformation is another important preprocessing step, particularly when using algorithms like K-Medoids, which require numerical inputs. Normalization, a common type of transformation, rescales features to a specific range or distribution, improving model performance and convergence speed. One widely used technique is Min-Max Scaling, which scales data to a fixed range, typically

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

[0,1]. In this method, the minimum value becomes 0, the maximum becomes 1, and all other values are scaled proportionally in between [29]. For categorical data, one-hot encoding is often used such as educational research [30]. This technique transforms a categorical variable with k possible values into k binary columns, each representing the presence or absence of a particular category. In this study, one-hot encoding will be applied to handle categorical variables. Effective preprocessing, especially using appropriate normalization and encoding techniques, is essential to achieving better predictive performance and reducing training time [31].

2.3. Grouping

This grouping process follows the concept of group investigation in cooperative learning, where group members have similar interests and different student characteristics [9], [10], [11]. The grouping process had three stages of grouping.

The first stage is interest-based grouping**.** In this stage, the initial dataset, consisting of 41 student records, was first divided based on their selected interests. Since interest is a binary attribute in this case (data warehouse development and performance measurement), this results in two separate groups: one consisting of data warehouse development and Performance measure.

The Second stage is homogenous grouping with K-Medoids. In this stage, K-Medoids clustering was independently applied to each interest-specific subset of the data. The K-Medoids algorithm, specifically the Partitioning Around Medoids (PAM) variant, was employed. Unlike K-Means, which defines cluster centers as centroids (which are the mean positions and may not correspond to actual data points), PAM selects actual data points as cluster centers (medoids). This makes PAM more robust to noise and outliers, as medoids are less sensitive to extreme values [32]. Common distance measures used in educational contexts include Euclidean, Manhattan, and Gower distance [15],[16], [33], [34], [35], [36], [37]. In this study, we try each different distance measure with K-Medoids and selected the best-performing distance measure based on the highest Silhouette Score, which ranges from -1 to +1, where higher values indicate that the data point is well-clustered.

The final stage is heterogeneous grouping. Based on the clustering results, student groups were then formed by selecting members alternately from different clusters, within each until each group contained 4 to 5 members. This approach ensured each group included heterogeneity student characteristics [15] and interest consistency within each group.

2.4. Evaluation

In the evaluation phase, group heterogeneity was quantified using Gower dissimilarity which is particularly well-suited for datasets containing mixed attribute types, such as numerical, ordinal, and categorical variables [34]. To better reflect real-world diversity, Gower dissimilarity was calculated using the original (preprocessed) dataset, leveraging its ability to handle missing values naturally. This approach provides a more realistic representation of student diversity within each group. Although Gower's metric is often used in clustering algorithms [38], it is equally valuable for exploratory and post-hoc statistical analysis, particularly when quantifying pairwise dissimilarity in mixed-type data [39]. In this study, Gower dissimilarity is functionally equivalent to Gower distance, and both terms refer to the same measure of how different two student profiles are. The formula dissimilarity between objects i and j is shown as formula:

$$d_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} d_{ijk}}{\sum_{k=1}^{p} W_{ijk}} \quad (1)$$

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

In formula, p is the number of attributes, w_{ijk} is a weight that equals 1 if both values for attribute k are present (and 0 otherwise), and d_{ijk} the dissimilarity between the values of objects i and j on attribute k. For numerical attributes, the dissimilarity is computed as:

$$d_{ij} = \frac{|x_{ik} - x_{jk}|}{R_k} \quad (2)$$

Where R_k is the range (max - min) of attribute k. For categorical attributes, $d_{ijk} = 0$ if $x_{ik} = x_{jk}$, and 1 if they differ. This metric enables a robust evaluation of group composition, where a lower average intra-group Gower dissimilarity indicates more homogeneous groups, and higher intra-group dissimilarity reflects more heterogenous.

3. RESULT

3.1. Data Preprocessing

To prepare the dataset for analysis, all relevant student profile from Table 1Figure *1*. data were integrated into a single dataset to support analysis, resulting in a unified dataset comprising all necessary attributes from various sources with result 11 column 41 row. Once this first step was completed, we saved for further analysis at section evaluation to raw data using Gower dissimilarity. In addition, it is also conveyed about the discussion of the research and testing that has been done.

Missing values were addressed using statistical strategies depending on the data type as refer on [27], [28]. For numeric columns such as GPA, Quiz 1, Quiz 2, and Quiz 3, missing entries were filled using the mean of the respective column. For categorical attributes like MBTI, VARK, Interest, BASDAT, DATA MINING, and Gender, missing values were filled using the mode (most frequent value) of each column.

In addition, the DWBI quiz performance was represented by a newly constructed variable named RataQuiz, calculated as the average score of the three quizzes (Quiz 1, Quiz 2, and Quiz 3) because we just need one quiz score and after it just delete the three quiz variables because we don't need them. This transformation was made to simplify and unify quiz-related performance into a single metric. Below is a sample of the result of preprocessing Table 2.

Table 2. Sample of Results Data Preprocessing

NIM	GPA	BASDAT	DATA MINING	MBTI	VAK	Gender	RataQuiz	Interest
	3.32	AB	AB	ENTJ	K	Female	7.33	Performance measurement
	3.23	A	BC	ISFP	A	Male	6.00	Performance measurement
	3.6	AB	A	ENTJ	V	Male	9.33	Data warehouse development
	3.47	A	A	ENTJ	A	Male	8.66	Performance measurement
	3.4	AB	AB	ENFP	V	Female	8.66	Performance measurement

3.2. Data Transformation

As on methodology on grouping process stage 2 using K-Medoids, we will try Gower, Manhattan, and Euclidean and will choose the best distance and k value that give the highest silhouette score. Since Euclidean and Manhattan distances only work with numeric data, all categorical features must be

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

encoded numerically. The classification of data types in this study refers to Table 3. On the other hand, Gower distance can handle a mix of numeric and categorical features, but to ensure proper treatment of ordinal features, the grades for "BASDAT" and "DATA MINING" are mapped to numbers reflecting their order of performance. The process begins by applying custom grade mapping, converting letter grades (e.g., A, AB, B) into corresponding ordinal values (1 to 9). This mapped data is then split and scaled in two different ways:

1. Gower Dataset Preparation:
   a. Numeric features are scaled using MinMaxScaler.
   b. Categorical features are kept in their original (raw) form.
   c. Gower no need PCA
2. Euclidean-Manhattan Dataset Preparation:
   a. Categorical features are one-hot encoded to convert them into numeric format.
   b. All features (including the newly one-hot encoded ones) are then scaled.
   c. PCA (Principal Component Analysis) is applied to reduce dimensionality, preparing the data for efficient distance-based computations.

Table 3. Variable Classification

Classification	Variable Name
Identifier	NIM
Interest	Student interest which is categorical
Numeric	GPA, RataQuiz
Categorical	MBTI, VAR, Gender
Ordinal	BADSAT, DATA MINING

This dual preparation ensures that each type of distance metric can be applied correctly according to its requirements, with the Gower setup handling mixed data types and the Euclidean/Manhattan setup working purely with numeric input.

3.3. Grouping

This study applied a three-stage grouping strategy, following to section 2.3 to form student groups based on interest and profile diversity. The grouping process aimed to generate intra-group heterogeneity while preserving intra-group interest alignment. The first grouping stage is interest-based grouping. In this stage we defied the data into subsets: those who interested in “Performance measure” and those in “Data warehouse development”. This split is applied to both of dataset -Gower dataset and Euclidean-Manhattan dataset. From this grouping of 41 records, 25 records belong to performance measure and 16 records to warehouse development.

The next stage is the homogenous grouping with K-Medoids PAM method. The Silhouette Score analysis was conducted to evaluate the clustering performance on two subsets: Data Warehouse Development and Performance Measurement. Three distance metrics were compared—Euclidean (with PCA), Manhattan (with PCA), and Gower distance—across varying numbers of clusters (k = 2 to 5). The following features were excluded from clustering: NIM (identifier) and interest (homogeneous within group).

For performance measurement subset as shown in Figure 2., Silhouette Score were computed for k = 2 to k = 5. The PCA + Euclidean and PCA + Manhattan methods performed best, both peak Silhouette Score at k = 4 with scores of approximately 0.66 and 0.625 respectively. This indicates that the optimal cluster structure for this subset is likely four clusters. In contrast, the Gower distance method showed low Silhouette Score across all k-values, with a peak below 0.08. This indicates poor separation, suggesting that PCA-based approaches more effectively captured the variance in student profiles.

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

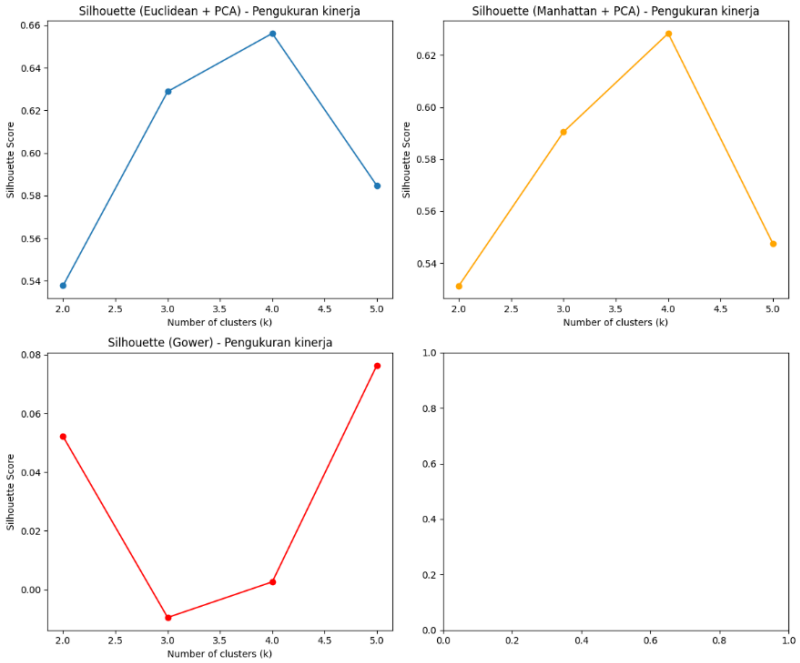

Figure 2. Silhouette Score Subset Performance Measurement

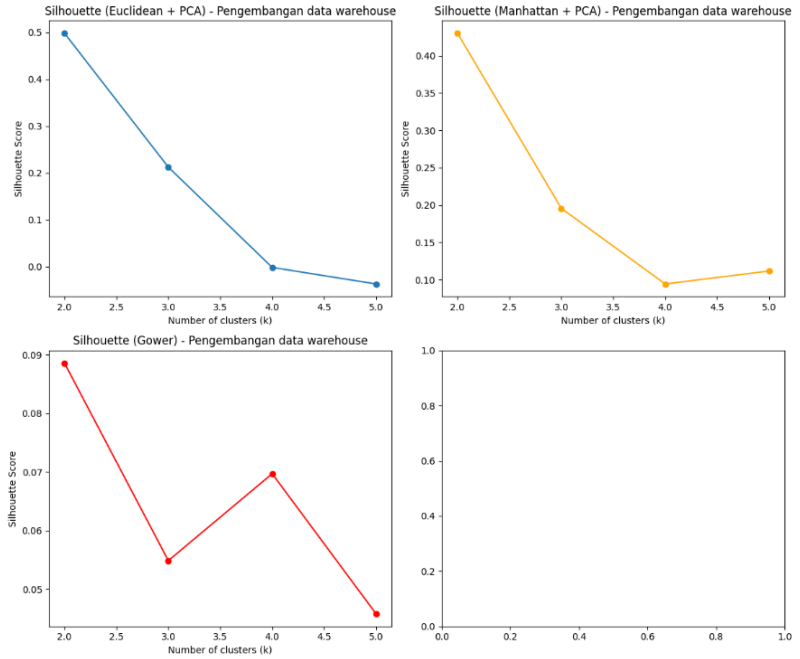

Figure 3. Silhouette Score Subset Data Warehouse Development

In the Data Warehouse Development subset, the Silhouette Score were generally lower. The best results were again achieved using Euclidean and Manhattan distances with PCA as shown on Figure 3., with optimal clustering found at k = 2, yielding Silhouette Score of around 0.50 and 0.42 respectively. This implies a more distinct and well-separated cluster structure with two clusters. Gower distance again showed poor performance, with Silhouette Score peaking at only ~0.09, indicating weak clustering validity.

In summary, the combination of PCA with Euclidean or Manhattan distance consistently outperformed Gower distance across both datasets. Since we aim to use a consistent distance metric for evaluating both datasets, and Euclidean distance produced the highest Silhouette Score overall, we

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

selected Euclidean distance with PCA as the most appropriate approach for clustering in this study. To visualize the clustering results using this method and the optimal number of clusters (k), Figure 4. illustrate the resulting cluster structures. The clustering results for the Performance Measurement subset produced four clusters (Cluster 0 to Cluster 3), with 7, 8, 6, and 4 students respectively. For the Data Warehouse Development subset, two clusters were identified: Cluster 0 with 7 students and Cluster 1 with 9 students.

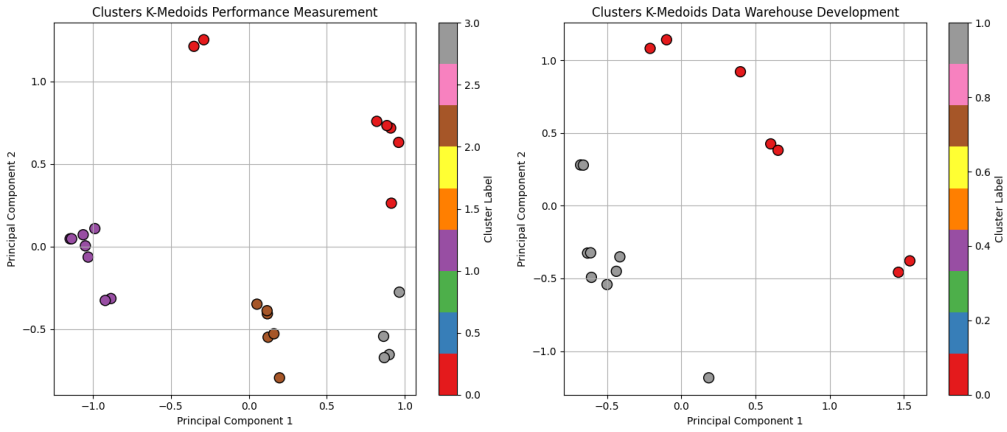

Figure 4. Cluster Visualization

The final stage of the grouping process was the formation of heterogeneous student groups, guided by the clustering outcomes from Stage 2. In this stage, students were assigned to groups by alternately selecting individuals from different clusters within the same domain of interest, maintaining group sizes of 4 to 5 members. Specifically, groups in the Performance Measurement subset were formed with up to 5 members resulting 5 Group (Group 1-5), while those in the Data Warehouse Development subset were limited to 4 members resulting 4 Group (Group 6-9). This strategy ensured each group exhibited diverse student characteristics [16] while preserving alignment in their academic interests. For better understanding, here is the visualization of grouping result on Figure 5

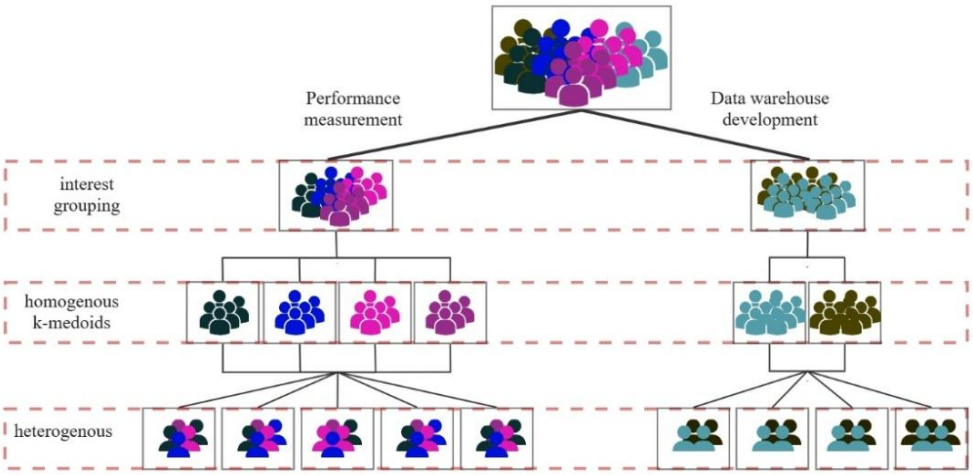

Figure 5. Stages of Grouping Process

3.4. Evaluation

To assess the heterogeneity within each final group, Gower dissimilarity was calculated using the original dataset. The analysis involved selecting the students from each final group (assembled during

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

Stage 3) and calculating the pairwise Gower dissimilarity among them. The mean of these distances served as a proxy for heterogeneity.

As per standard Gower dissimilarity computation, missing values are ignored during pairwise attribute comparisons, meaning only attributes present in both records are used. This approach allows flexible handling of incomplete data but may introduce slight imprecision or especially when the number of comparable attributes varies significantly between pairs. When too few attributes are available, such as in the case of a student with almost all missing values, the distance cannot be computed, resulting in a NaN. This occurred in Group 1 due to one such student. However, if at least some attributes are shared, the distance can still be calculated based on the available data.

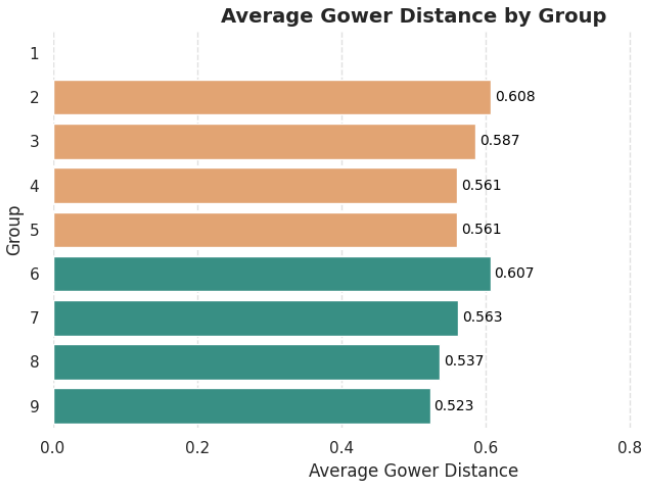

Figure 6. Average Gower dissimilarity by Group

As shown in Figure 6., the average intra-group Gower dissimilarity for Performance Measurement subset (Group 1 – Group 5) ranged from 0.608 in Group 2 to 0.561 in Group 4-5 indicating varying levels of profile heterogeneity across teams. For the Data Warehouse Development subset (Groups 6–9), the average intra-group distance ranged from 0.607 in Group 6 to 0.523 in Group 9. Higher average Gower dissimilarity reflects greater diversity, which is generally desirable for fostering richer collaboration in group-based investigative learning. One exception was Group 1, where the presence of a student with extensive missing data led to undefined pairwise distances in some comparisons.

4. DISCUSSIONS

In grouping stage 2 (homogenous grouping), the application of K-Medoids clustering with Euclidean distance resulted in higher silhouette score and clearer cluster separation. This outcome aligns with findings by Keefe Murphy et al. [35], who reported strong performance of K-Medoids clustering using Euclidean distance in educational datasets. Notably, our results demonstrate that PCA contributed not only to dimensionality reduction but also to noise suppression across mixed categorical-numerical attributes, improving Silhouette Score compared to Gower distance. However, the superior clustering performance of Euclidean distance after PCA suggests that well-scaled continuous representations may outweigh the theoretical flexibility of Gower (which designed for mixed data) in certain educational datasets.

However, in Stage 3 (final group formation), where the goal was to form heterogeneous groups by combining students from different clusters, an interesting pattern emerged. In both subsets (Performance Measurement and Data Warehouse Development), the average Gower distance—used as

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

a proxy for heterogeneity—decreased in later-numbered groups (e.g., Group 5), suggesting those groups were more homogeneous.

This pattern is likely due to unequal cluster sizes produced during the clustering process. For example, Cluster 1, being larger, contributed more students to the final groupings. As students were drawn sequentially from each cluster to form the groups, the earlier groups (e.g., Group 1) maintained more balanced cluster representation, while later groups (e.g., Group 5) became increasingly dominated by students from the larger cluster. This sequential assembly process unintentionally reduced diversity in those later groups.

This unequal distribution of student characteristics across groups may affect overall group quality, particularly in terms of balance, role distribution, and inter-member learning potential. Previous research has shown that such disparities can influence group dynamics and learning outcomes [40], [41].

From a pedagogical perspective, this has implications for cooperative learning strategies, such as Group Investigation, where balanced group diversity is considered essential. To mitigate these limitations in future work, alternative group assembly strategies could be explored—such as constraining maximum cluster contribution or using dynamic selection algorithms that adapt to cluster sizes. While our current method was simple and interpretable, future iterations would benefit from enhancements that better preserve the intended heterogeneity of the learning groups

5. CONCLUSION

This study developed and validated a data-driven methodology for forming heterogeneous student groups tailored to the Group Investigation cooperative learning model. By integrating student profile data (including cognitive traits and academic interests) and applying PCA enhanced clustering with Euclidean distance, the approach successfully identified homogenous clusters which were then reorganized into interest-aligned yet diverse groups. Evaluation using Gower dissimilarity on the raw dataset confirmed that the method generally achieved internal heterogeneity which is an important factor in enhancing collaborative learning potential.

Notably, the analysis revealed that imbalances in cluster size can subtly influence the distribution of diversity across final groups, with later-formed groups tending toward increased homogeneity. This insight underscores the importance of incorporating cluster balancing strategies into future group assembly designs.

Overall, the findings affirm that clustering-based group formation is a promising approach to support heterogeneity in collaborative learning. Future research is encouraged to extend this method across broader, more varied student populations, and to empirically investigate its pedagogical impact on student engagement, performance, and group dynamics.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] M. Malan, “The Effectiveness of Cooperative Learning in an Online Learning Environment Through a Comparison of Group and Individual Marks,” *Electronic Journal of e-Learning*, vol. 19, no. 6, pp. 588–600, Dec. 2021, doi: 10.34190/ejel.19.6.2238.

[2] A. Abramczyk and S. Jurkowski, “Cooperative learning as an evidence-based teaching strategy: what teachers know, believe, and how they use it,” *Journal of Education for Teaching*, vol. 46, no. 3, pp. 296–308, 2020, doi: 10.1080/02607476.2020.1733402.

[3] Sugiharto, “Geographical students’ learning outcomes on basic political science by using cooperative learning model with Group Investigation (GI) type in State University of Medan,

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

Indonesia," *J Hum Behav Soc Environ*, vol. 30, no. 4, pp. 447–456, May 2020, doi: 10.1080/10911359.2019.1696261.

[4] J. Kilpeläinen-Pettersson, P. Koskinen, A. Lehtinen, and T. Mäntylä, "Cooperative Learning in Higher Education Physics – A Systematic Literature Review," *Int J Sci Math Educ*, Jan. 2025, doi: 10.1007/s10763-024-10538-3.

[5] M. Susanti, S. Suyanto, J. Jailani, and H. Retnawati, "Problem-based learning for improving problem-solving and critical thinking skills: A case on probability theory course," *Journal of Education and Learning (EduLearn)*, vol. 17, no. 4, pp. 507–525, Nov. 2023, doi: 10.11591/edulearn.v17i4.20866.

[6] M. Saqr and S. López-Pernas, "The temporal dynamics of online problem-based learning: Why and when sequence matters," *Int J Comput Support Collab Learn*, vol. 18, no. 1, pp. 11–37, 2023, doi: 10.1007/s11412-023-09385-1.

[7] D. A. Donovan and G. L. Connell, "Evolution of a Student-Centered Biology Class: How Systematically Testing Aspects of Class Structure Has Informed Our Teaching," in *Active Learning in College Science*, Cham: Springer International Publishing, 2020, pp. 307–323. doi: 10.1007/978-3-030-33600-4_20.

[8] M. Fischer, R. M. Rilke, and B. B. Yurtoglu, "When, and why, do teams benefit from self-selection?," *Exp Econ*, vol. 26, no. 4, pp. 749–774, Sep. 2023, doi: 10.1007/s10683-023-09800-2.

[9] Y. Sharan and S. Sharan, "Design for Change: A Teacher Education Project for Cooperative Learning and Group Investigation in Israel," in *Pioneering Perspectives in Cooperative Learning*, Routledge, 2021, pp. 165–182. doi: 10.4324/9781003106760-8.

[10] A. C. Seherrie and A. S. Mawela, "Life Orientation teachers' pedagogical content knowledge and skills in using a group investigation cooperative teaching approach," *Journal of Education*, no. 89, pp. 1–20, Jan. 2023, doi: 10.17159/2520-9868/i89a03.

[11] H. Silva, J. Lopes, E. Morais, and C. Dominguez, "Fostering Critical and Creative Thinking through the Cooperative Learning Jigsaw and Group Investigation," *International Journal of Instruction*, vol. 16, no. 3, pp. 261–282, Jul. 2023, doi: 10.29333/iji.2023.16315a.

[12] T. Vallès-Català and R. Palau, "Minimum entropy collaborative groupings: A tool for an automatic heterogeneous learning group formation," *PLoS One*, vol. 18, no. 3 March, 2023, doi: 10.1371/journal.pone.0280604.

[13] S. Lorente, M. Arnal-Palacián, and M. Paredes-Velasco, "Effectiveness of cooperative, collaborative, and interdisciplinary learning guided by software development in Spanish universities," *European Journal of Psychology of Education*, vol. 39, no. 4, pp. 4467–4491, Dec. 2024, doi: 10.1007/s10212-024-00881-y.

[14] R. C.-Y. Loh and C.-S. Ang, "Unravelling Cooperative Learning in Higher Education," *Research in Social Sciences and Technology*, vol. 5, no. 2, pp. 22–39, May 2020, doi: 10.46303/ressat.05.02.2.

[15] Kanika, S. Chakraverty, P. Chakraborty, and M. Madan, "Effect of different grouping arrangements on students' achievement and experience in collaborative learning environment," *Interactive Learning Environments*, vol. 31, no. 10, pp. 6366–6378, 2023, doi: 10.1080/10494820.2022.2036764.

[16] G. Nalli, D. Amendola, A. Perali, and L. Mostarda, "Comparative analysis of clustering algorithms and moodle plugin for creation of student heterogeneous groups in online university courses," *Applied Sciences (Switzerland)*, vol. 11, no. 13, 2021, doi: 10.3390/app11135800.

[17] G. Nalli, D. Amendola, and S. Smith, "Artificial Intelligence to Improve Learning Outcomes Through Online Collaborative Activities," in *Proceedings of the European Conference on e-Learning, ECEL*, Oct. 2022, pp. 475–479. doi: 10.34190/ecel.21.1.661.

[18] G. Nalli and S. Smith, "Comparison of the Effectiveness and Performance of Student Workgroups in Online Wiki Activities with and without AI †," *Engineering Proceedings*, vol. 56, no. 1, 2023, doi: 10.3390/ASEC2023-16273.

[19] Z. Bousalem, A. Qazdar, and I. El Guabassi, "Cooperative Learning Groups: A New Approach Based on Students' Performance Prediction," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 12, pp. 34–48, Aug. 2023, doi: 10.3991/ijoe.v19i12.41181.

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

[20] N. Gavrilovic, T. Sibalija, and D. Domazet, "Design and implementation of discrete Jaya and discrete PSO algorithms for automatic collaborative learning group composition in an e-learning system," *Appl Soft Comput*, vol. 129, p. 109611, Nov. 2022, doi: 10.1016/J.ASOC.2022.109611.

[21] O. S. Revelo, C. A. Collazos Ordonez, M. A. Redondo, and I. Ibert Bittencourt Santana Pinto, "Homogeneous Group Formation in Collaborative Learning Scenarios: An Approach Based on Personality Traits and Genetic Algorithms," *IEEE Transactions on Learning Technologies*, vol. 14, no. 4, pp. 486–499, 2021, doi: 10.1109/TLT.2021.3105008.

[22] P. S. Pawar, J. R. Saini, P. J. Pawar, and S. Vaidya, *Genetic Algorithm Application for Improving the Performance of Teaching Learning Process Through Collaborative Learning*, vol. 1086 LNNS. 2024. doi: 10.1007/978-981-97-6036-7_27.

[23] C.-C. Peng, C.-J. Tsai, T.-Y. Chang, J.-Y. Yeh, and M.-C. Lee, "Novel heterogeneous grouping method based on magic square," *Inf Sci (N Y)*, vol. 517, pp. 340–360, 2020, doi: 10.1016/j.ins.2019.12.088.

[24] E. Schubert and P. J. Rousseeuw, "Fast and eager K-Medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms," *Inf Syst*, vol. 101, p. 101804, Nov. 2021, doi: 10.1016/J.IS.2021.101804.

[25] J. Wang, X. W. Chen, K. R. Cheng, Y. L. Cao, and B. Pan, *A Study on the Characteristics of College Students' Consumption Behavior Based on Clustering and Association Rules*, vol. 1423. 2021. doi: 10.1007/978-3-030-78618-2_30.

[26] T. Hamim, F. Benabbou, and N. Sael, "Toward a Generic Student Profile Model," in *Innovations in Smart Cities Applications*, Edition 3., Springer, 2020, pp. 200–214. doi: 10.1007/978-3-030-37629-1_16.

[27] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00516-9.

[28] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artif Intell Rev*, vol. 53, no. 2, pp. 1487–1509, 2020, doi: 10.1007/s10462-019-09709-4.

[29] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," in *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 2020, pp. 729–735. doi: 10.1109/ICSSIT48917.2020.9214160.

[30] A. Tran, C. Zuniga-Navarrete, L. J. Segura, A. Dourado, X. Wang, and C. R. Bego, "Categorical Variable Coding for Machine Learning in Engineering Education," in *Proceedings - Frontiers in Education Conference, FIE*, 2024. doi: 10.1109/FIE61694.2024.10893080.

[31] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies (Basel)*, vol. 9, no. 3, 2021, doi: 10.3390/technologies9030052.

[32] M. Prince and P. M. Joe Prathap, "A Novel Approach to Design Distribution Preserving Framework for Big Data," *Intelligent Automation and Soft Computing*, vol. 35, no. 3, pp. 2789–2803, 2023, doi: 10.32604/iasc.2023.029533.

[33] U. Mishra and S. S. Parikh, "Clustering method for forming student groups based on the cognitive performance measured using eye responses," in *Proceedings of 2024 3rd International Conference on Artificial Intelligence and Intelligent Information Processing, AIIIP 2024*, 2025, pp. 303–310. doi: 10.1145/3707292.3707381.

[34] K. Murphy, S. López-Pernas, and M. Saqr, *Dissimilarity-Based Cluster Analysis of Educational Data: A Comparative Tutorial Using R*. 2024. doi: 10.1007/978-3-031-54464-4_8.

[35] T. J. Fontalvo-Herrera, E. J. Delahoz, and A. A. Mendoza-Mendoza, "Application of data mining for the classification of university programs of industrial engineering accredited in high quality in Colombia | Aplicación de minería de datos para la clasificación de programas universitarios de ingeniería industrial acreditad," *Informacion Tecnologica*, vol. 29, no. 3, pp. 89–96, 2018, doi: 10.4067/S0718-07642018000300089.

Vol. 6, No. 5, October 2025, Page. 3886-3898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4765

[36] Q. Hu and H. Rangwala, "Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students," in *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020*, 2020, pp. 431–437.

[37] D. Ebbert and S. Dutke, "Patterns in students' usage of lecture recordings: A cluster analysis of self-report data," *Research in Learning Technology*, vol. 28, 2020, doi: 10.25304/rlt.v28.2258.

[38] P. Liu, H. Yuan, Y. Ning, B. Chakraborty, N. Liu, and M. A. Peres, "A modified and weighted Gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses," *BMC Med Res Methodol*, vol. 24, no. 1, 2024, doi: 10.1186/s12874-024-02427-8.

[39] J. De La Cruz Hernandez, K. J. Tobin, J. C. Kilburn, and M. E. Bennett, "Using a Modified Gower Distance Measure to Assess Supplemental Learning Supporting an Online Social Science Graduate Course," *Educ Sci (Basel)*, vol. 15, no. 3, 2025, doi: 10.3390/educsci15030371.

[40] C. Alalouch, "Cognitive styles, gender, and student academic performance in engineering education," *Educ Sci (Basel)*, vol. 11, no. 9, 2021, doi: 10.3390/educsci11090502.

[41] S. Jiang, J. McClure, C. Tatar, F. Bickel, C. P. Rosé, and J. Chao, "Towards inclusivity in AI: A comparative study of cognitive engagement between marginalized female students and peers," *British Journal of Educational Technology*, vol. 55, no. 6, pp. 2557–2573, 2024, doi: 10.1111/bjet.13467.