DOI: 10.14421/ijid.2024.4632 IJID (International Journal on Informatics for Development), *e-ISSN: 2549-7448*
Vol. 13, No. 2, December 2024, Pp. 460-472

K-Means Clustering of Social Studies Performance at Junior High School

Tundo*
Department of Information Systems
Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika
Jakarta, Indonesia
asna8mujahid@gmail.com

Syifa Raihanah*
Department of Information Systems
Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika
Jakarta, Indonesia
syifaraihanah48@gmail.com

Tri Wahyudi
Department of Information Systems
Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika
Jakarta, Indonesi
triwahyudi199003@gmail.com

Sugiyono
Department of Information Systems
Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika
Jakarta, Indonesia
inosoguy007@gmail.com

Article History
Received July 6th, 2024
Revised September 1st, 2024
Accepted October 7th, 2024
Published December, 2024

***Abstract*— This study aims to optimize the use of technology in evaluating student performance by grouping students based on their abilities. The main issues include the underutilization of technology, the absence of an appropriate evaluation system for different levels of student ability, and ineffective methods for grouping students. The K-Means Clustering algorithm was chosen because it has proven effective in grouping academic data in various studies. The data used includes Daily Knowledge Scores (DKS), Daily skill scores (DSS), Mid-term Summative Scores (MSS), End-of-Year Summative Scores (ESS), and Grade Report (GR). The data was analyzed using the CRISP-DM methodology with the help of RapidMiner. The results showed that 28.63% of students were classified as having excellent performance, 50.21% as having good performance, and 21.16% as having moderate performance. The Davies-Bouldin Index score of 1.713 for K=3 was considered sufficient for distinguishing the different student performance groups. The results of this study are expected to help schools provide learning support that better aligns with student needs. Future research is recommended to focus on optimizing the number of clusters (K), applying this method to other subjects, and integrating it with e-learning platforms for real-time student performance monitoring.**

Keywords—academic data; e-learning; Rapidminer; student ability; student performance

1 INTRODUCTION

In the current digital era [1], education faces challenges in fully utilizing technology, as seen at Ksatrya Junior High School, Jakarta, which has a student grading system that has not yet been fully leveraged to support learning and performance evaluation [2]. The existing performance evaluation system has not been able to give special attention to students with varying academic abilities [3], making it difficult for teachers to provide appropriate interventions based on students' needs.

As the number of students increases [4], the importance of more accurate student performance mapping becomes more evident. Without proper mapping, students with average achievements do not receive adequate support [5], while high-performing students are not sufficiently challenged in the learning process. At Ksatrya Junior High School, Jakarta, there is no effective method for grouping students based on their performance in Social Studies, making the learning process less than optimal [6].

In addressing this issue, various approaches in machine learning can be applied for data analysis, including supervised and unsupervised learning methods. This research focuses on unsupervised learning, which allows for data clustering without labels [7]. The most commonly used unsupervised learning methods include hierarchical clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and K-Means Clustering [8].

Hierarchical clustering reveals the hierarchical structure of complex data but performs slowly on large datasets, making it less suitable for this research. Meanwhile [9], DBSCAN is effective in identifying irregular clusters and handling noise, but it is challenging to determine parameters such as 'epsilon' and is less suitable for data with small density variations, as seen in student performance in this study [10].

K-means clustering is an effective and simple method for grouping large amounts of data based on similarities between the data points [11]. This method excels in terms of speed and computational efficiency, especially when dealing with large datasets and round-shaped clusters. However, K-Means Clustering has some drawbacks, such as being sensitive to outliers and only being able to cluster linear-shaped data. Despite these limitations, K-Means Clustering is often used due to its simplicity and relatively accurate results in many scenarios.

K-means clustering has also been widely applied in various previous studies. For example, in the journal [12] the authors faced challenges in the selection process for BETUNAS scholarship recipients. This process often encountered difficulties in managing large datasets and both academic and non-academic variables, leading to a lack of objectivity and inefficiency. To address these issues, the authors applied the K-Means Clustering algorithm to group students based on their academic and non-academic achievements and developed a decision support system to enhance the objectivity and efficiency of the selection process. The study aimed to create a more transparent and data-driven scholarship selection process. The results showed that out of 200 BETUNAS student records, 43 students in Cluster 0 were eligible for the scholarship, while 157 students in Cluster 1 were not eligible. In conclusion, the K-Means Clustering algorithm successfully facilitated the objective and efficient grouping of scholarship recipients, ensuring that the selected students truly deserved based on their performance.

In the journal [13], it was found that the main issue was the absence of a system to assist students in selecting a concentration within the Informatics program, which includes Software Engineering, Network Engineering, and Data Analytics. This often led to a mismatch between students' concentration choices and their academic abilities. To address this issue, the authors developed a recommendation system utilizing the K-Means algorithm, employing academic data from students in semesters 1 to 4 to group them based on their academic abilities. This study aimed to provide improved guidance for students in selecting the appropriate concentration, thereby minimizing errors in their specialization choices. The implementation results showed that the recommendation system achieved an accuracy of 81%, significantly higher than the previous recommendation system, which had an accuracy of only 7.55%. In conclusion, the K-Means method successfully provided more accurate recommendations in assisting students to choose concentrations that align with their academic abilities, offering an alternative for the Informatics program at Bina Darma University to provide more precise academic guidance.

In the research [14], the authors developed a web-based information system and application to address the issue of unclear classification and the absence of clear criteria for advanced classes at SMK Al-Badar Balaraja. This application utilizes the K-Means Clustering method to group students into five clusters: Class A, B, C, D, and E, based on criteria such as report card grades, attendance, achievements, extracurricular activities, and involvement in student organizations. Out of the 164 students analyzed, the clustering results were distributed as follows: 34 students in Class A, 28 in Class B, 35 in Class C, 27 in Class D, and 40 in Class E. Testing of this application demonstrated that the system is effective and reliable in identifying students eligible for placement in advanced classes, which is expected to improve the quality of education at SMK Al-Badar Balaraja. This research focused on 10th-grade students in the Office Administration major at SMK Al-Badar Balaraja, located in the Balaraja District, Tangerang Regency.

Various studies have demonstrated that the K-Means Clustering algorithm effectively groups student academic data for various purposes, such as scholarship selection, study concentration choices, and class classification. Based on its proven success, this study utilizes K-Means Clustering due to its advantages in grouping student performance patterns that are clear and easily identifiable. This research aims to develop a student performance clustering method in Social Studies [15], with the expectation of identifying clearer and more structured performance patterns. The results of this clustering will serve as the foundation for providing recommendations to the school, aiding in strategic decision-making [16] to offer more tailored interventions based on each student's academic needs and abilities.

2 METHOD

In this study, the [17] CRISP-DM (Cross Industry Standard Process for Data Mining) methodology will be employed to manage and process the data collected. CRISP-DM was chosen due to its systematic yet flexible approach [18], making it well-suited for the data clustering process in this research. The following are the stages of the CRISP-DM methodology and a flowchart as shown in Fig. 1.

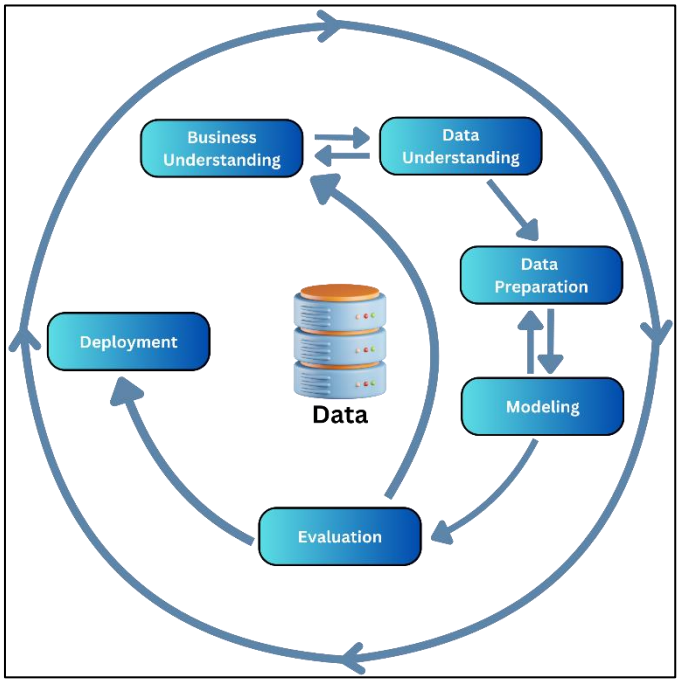

Figure 1. CRISP-DM

2.1 Business Understanding

In the [19] business understanding phase, an observation was conducted on May 14, 2024, from 09:00 to 09:45 AM WIB, followed by an interview with the social studies teacher and the student grade management system administrator from 09:46 to 10:30 AM WIB, along with a literature review to identify the requirements for analyzing student performance in social studies. Based on the results of the observation, interview, and literature review, it was found that grouping student performance is necessary to understand existing patterns. Therefore, the K-Means Clustering algorithm was selected for its ability to categorize student data into several clusters, which is expected to provide a more structured overview of student performance.

2.2 Data Understanding

In this phase, performance data for 241 students was collected from the student grading system at Ksatrya Junior High School, Jakarta. The student performance [20] data comprises 16 attributes, including student ID, NISN (National Student Identification Number), student name, Daily Knowledge Scores (DKS) ranging from 1 to 4, Daily Skill Scores (DSS) ranging from 1 to 4, Mid-term Summative Scores (MSS), End-of-Year Summative Scores(ESS), Grade Report (GR), attitude scores, and descriptions. RapidMiner was utilized to explore the data, including checks for feature types, missing values, and descriptive statistics such as maximum values and averages [21]. The results of the data exploration are presented in Table 1.

Table 1. Data Description

Name	Type	Missing Value	Max	Average /Values
ID	Integer	0	241	120.632
NISN	Integer	0	9390	9251.756
Name	Nominal	0	124 Student	All students names
DKS[a] 1	Integer	0	100	89.565
DKS 2	Integer	0	100	86.196
DKS 3	Integer	0	100	91.464
DKS 4	Integer	0	100	88.120
DSS[b] 1	Integer	0	100	89.201
DSS 2	Integer	0	100	89.455
DSS 3	Integer	0	100	92.134
DSS 4	Integer	0	100	87.995
MSS[c]	Integer	0	100	83.708
ESS[d]	Integer	0	100	83.254
GR[e]	Integer	0	100	86.407
Attitude Score	Nominal	0	A	All attitude values
Descriptions	Nominal	0	8 Student Description	All student descriptions

a. DKS: Daily Knowledge Scores
b. DSS: Daily Skill Scores
c. MSS: Mid-term Summative Scores
d. ESS: End-of-Year Summative Scores
e. GR: Grade Report

2.3 Data Preparation

During the previous phase, attributes such as Student ID, NISN, Student Name, Attitude Score, and Description were identified as nominal or integer types. In the [22] data preparation phase, the Select Attributes operator in RapidMiner was employed to select relevant integer-type attributes for the K-Means algorithm. Table 2 presents the selected attributes relevant to the analysis objectives.

After the selection process, data normalization was applied using the Normalize operator to standardize the value ranges across attributes, ensuring that differences in scale do not introduce bias during the clustering process [23]. Formula 1 presents the Z-transformation normalization formula that was utilized.

Table 2. Attribute Selection Results

ID[a]	DKS			DSS			MSS	ESS	GR
	1	…	*4*	*1*	…	*4*			
1	100	…	85	90	…	90	91	97	94
2	70	…	85	100	…	90	80	71	84
3	85	…	85	100	…	90	91	72	87
4	95	…	85	100	…	100	90	81	92
5	93	…	90	100	…	85	87	92	93
6	93	…	85	80	…	80	89	89	87
7	75	…	100	80	…	80	100	84	88
8	100	…	90	90	…	90	100	83	92
9	80	…	80	80	…	80	80	76	80
10	70	…	90	80	…	80	100	80	86
…	…	…	…	…	…	…	…	…	…
241	90	…	90	100	…	90	90	80	88

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

The Z-transformation formula was used for data normalization, where Z represents the transformed result, μ is the mean, and σ is the standard deviation of the data [24]. Table 3 presents the normalized data results.

Table 3. Normalized Data Results

ID	DKS			DSS			MSS	ESS	GR
	1	…	*4*	*1*	…	*4*			
1	0.9	…	-0.4	0.1	…	0.3	0.9	1.7	2.0
2	-1.6	…	-0.4	1.6	…	0.3	-0.4	-1.5	-0.6
3	-0.4	…	-0.4	1.6	…	0.3	0.9	-1.3	0.2
4	0.4	…	-0.4	1.6	…	1.9	0.8	-0.3	1.5
5	0.3	…	0.3	1.6	…	-0.5	0.4	1.1	1.7
6	0.3	…	-0.4	-1.3	…	-1.3	0.6	0.7	0.2
7	-1.2	…	1.8	-1.3	…	-1.3	2.0	0.1	0.4
8	0.9	…	0.3	0.1	…	0.3	2.0	0.0	1.5
9	-0.8	…	-1.2	-1.3	…	-1.3	-0.4	-0.9	-1.6
…	…	…	…	…	…	…	…	…	…
241	0.0	…	1.8	0.1	…	0.3	0.3	-0.4	0.4

2.4 *Modeling*

In the modeling phase, the K-Means Clustering algorithm was applied to group student data based on similarities in their scores. The K-Means operator in RapidMiner was used to execute this process [25]. The following are the steps in the K-Means modeling process, illustrated in Fig. 2.

2.4.1 *Determining the Number of Clusters:* The number of clusters used was set to 3, following the grading system outlined in the school curriculum [26], which classifies student performance into three categories: Excellent, Good, and Moderate.

2.4.2 *Determining Initial Cluster Centroids:* Once the number of clusters was determined, the algorithm randomly selected initial centroids for each cluster. These centroids serve as the initial reference points for the clustering process [27]. Manually, the centroids were selected from the objects representing the minimum value, the value closest to the mean, and the maximum value from the entire dataset. Table 4 below presents the initial normalized centroids chosen manually based on these criteria.

2.4.3 *Calculating Distance to Each Centroid:* Each student's data was measured against the centroids using the Euclidean distance metric [28]. Formula 2 presents the Euclidean distance formula used to calculate the distance between data points and centroids.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + \cdots + (x_n - y_n)^2} \quad (2)$$

The distance calculated between data points and centroids using attributes x_1, y_1 for the data points and x_2, y_2 for the centroids. The result is the distance (d) between the data and the centroid. With 241 data points, it is impractical to manually display all calculations. As an illustration, a manual calculation of the distance between student objects with ID 1 and the centroid is provided.

Table 4. Initial Centroid Results

ID	Centroid	DKS			DSS			MSS	ESS	GR
		1	…	*4*	*1*	…	*4*			
46	1	0.8	…	1.8	-1.5	…	1.9	1.9	0.8	2.7
12	2	-5.7	…	-1.1	-1.3	…	-1.2	-0.9	-1.5	-3.3
61	3	-0.3	…	-0.4	0	…	0.3	0.5	-0.8	-0.3

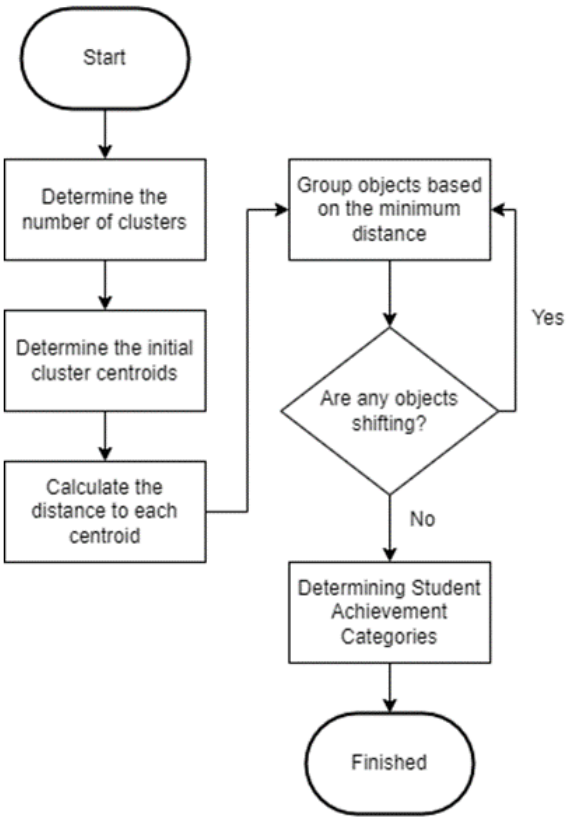

Figure 2. Flowchart of K-Means

$$C_1 = \sqrt{(0.9-(-0.9))^2 + (2.4-2.4)^2 + (1.2-(-0.2)^2 + ((-0.4)-1.8)^2 + (0.1-1.6)^2 + (0.1-1.5)^2 + ((-0.3)-1.3)^2 + (0.3-1.9)^2 + (0.9-2.0)^2 + (1.7-0.8)^2 + (2.0-2.7)^2} = 4.3$$

$$C_2 = \sqrt{(0.9-(-5.7))^2 + (2.4-(-1.0)^2 + (1.2-(-0.2))^2 + ((-0.4)-(-1.2))^2 + (0.1-(-1.3)^2 + (0.1-(-1.3))^2 + ((-0.3)-(-1.9))^2 + (0.3-(-1.3))^2 + (0.9-(-0.9))^2 + (1.7-(-1.6))^2 + (2.0-(-3.4))^2} = 10.4$$

$$C_3 = \sqrt{(0.9-(-0.4))^2 + (2.4-0.7)^2 + (1.2-(-0.9)^2 + ((-0.4)-(-0.4))^2 + (0.1-0.1)^2 + (0.1-0.1)^2 + ((-0.3)-(-0.3))^2 + (0.3-0.3)^2 + (0.9-0.5)^2 + (1.7-(-0.9))^2 + (2.0-0.3)^2} = 4.5$$

Subsequently, the distance calculation was continued for the second, third, and up to the 241st data point [29]. The results of the distance calculations between data and centroids are presented in Table 5.

2.4.4 *Assigning Objects Based on Minimum Distance:* Students were grouped into the cluster with the nearest centroid. The calculated distances were compared, and the student data was assigned to the cluster with the closest centroid. This distance indicates that the data point belongs to the group most closely aligned with the [30] nearest centroid. The results of the object grouping are presented in Table 6. Based on the preliminary clustering results, the number of objects in each cluster was 26 in Cluster 1, 3 in Cluster 2, and 212 in Cluster 3. The list of object IDs belonging to each cluster is shown below.

Cluster 1: 1, 4, 5, 39, 40, 41, 44, 46, 48, 49, 63, 66, 126, 155, 157, 170, 184, 189, 213, 217, 226, 227, 228, 229, 232, 233.

Cluster 2: 12, 13, 27.

Cluster 3: 2, 3, 6, 7, 8, 9, 10, 11, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 42, 43, 45, 47, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 156, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 185, 186, 187, 188, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 214, 215, 216, 218, 219, 220, 221, 222, 223, 224, 225, 230, 231, 234, 235, 236, 237, 238, 239, 240, 241.

Table 5. Euclidean Distance Results

ID	C1	C2	C3
1	4.308091813	10.42944753	4.52688347
2	6.325805518	7.780600183	3.771990459
3	5.280335973	8.598502793	3.147651055
4	3.37278322	10.79441306	4.539902023
5	3.708536393	10.64772502	4.82650929
6	7.973486623	7.672911102	4.414735581
7	7.742672871	7.556651423	5.064690167
8	4.301263219	9.675916254	3.519537981
9	9.433507249	5.493448697	4.429086723
10	8.21981049	6.466800443	4.56122063
…	…	…	…
…	…	…	…
241	4.787414716	8.499894959	2.899759446

Table 6. Object Grouping Results

ID	Nama	C1	C2	C3
1	Siswa_1	1		
2	Siswa_2			1
3	Siswa_3			1
4	Siswa_4	1		
5	Siswa_5	1		
6	Siswa_6			1
7	Siswa_7			1
8	Siswa_8			1
9	Siswa_9			1
10	Siswa_10			1
…	…	…	…	…
241	Siswa_241			1

2.4.5 *Checking Object Movements Between Clusters:* The algorithm checks whether any students moved from one cluster to another based on updated distance calculations. If movement is detected, the centroids are updated. Table 7 shows an example of the updated centroids in the second iteration. This iterative process continues until the clusters reach a stable or convergent state. Convergence in this context means that no further movements of objects between clusters occur, indicating that the centroid positions are fixed, and the clusters have been optimally formed.

2.4.6 *Student Performance Clustering:* Once the clusters stabilized, student data was successfully grouped into clusters that reflected their performance patterns. This clustering result was then analyzed further to understand the distribution and characteristics of student performance in Social Studies.

2.5 Evaluation

At this stage, the K-Means model is evaluated to assess the clustering quality. The evaluation is carried out using the Davies-Bouldin Index, implemented through RapidMiner. The Loop Parameter operator is used to test various cluster configurations, where each iteration is evaluated using the Davies-Bouldin Index. The purpose of this evaluation is to ensure that the model forms optimal clusters and that the chosen parameters result in efficient cluster separation.

2.6 Deployment

The results of the K-Means clustering of student performance data are used to recommend learning strategies at Ksatrya Junior High School, Jakarta. Each student group assists teachers in designing interventions tailored to their academic abilities. Students with moderate performance will receive appropriate support while excellent-performing students will be given more relevant challenges. The goal is to maximize student performance data to make learning more focused and personalized.

3 RESULT AND DISCUSSION

This section presents the results of the K-Means clustering analysis on student performance data and its evaluation, carried out using RapidMiner. Based on these results, recommendations are provided to improve student learning strategies.

3.1 K-Means Results

This section presents the results of the K-Means clustering process in the form of tables and graphical visualizations for ease of understanding. These results were produced using the RapidMiner application, which grouped student performance data into several clusters. The tables provide details for each cluster, while the graphical visualizations illustrate the patterns of each student group based on the analysis results.

3.1.1 *Cluster Centroid Results:* The K-Means clustering analysis yields centroids representing each attribute's average within the clusters. These centroids provide an overview of the key characteristics of each student group, based on their academic performance patterns. Table 8 presents the centroid values for each attribute in the three clusters. The centroid analysis shows that Cluster_0 has the highest total centroid value compared to the other clusters. This indicates that, overall, the attributes in this cluster have higher values, reflecting the dominant characteristics of the students within it. On the other hand, Cluster_2 has more balanced centroid values, with the attribute values tending to be stable and not too extreme. This reflects a more neutral characteristic in the students grouped within this cluster. Conversely, Cluster_1 has the lowest total centroid value, indicating that overall, the attribute values within this cluster are lower than in the other clusters

Table 7. New Centroid Results

Centroid	DKS			DSS			MSS	ESS	GR
	1	…	*4*	*1*	…	*4*			
1	0.8	…	1.2	1.2	…	1.3	1.0	0.2	1.6
2	-5.0	…	-0.7	-0.3	…	-0.8	-0.9	-0.2	-2.2
3	0.0	…	-0.1	-0.1	…	-0.2	-0.1	0.0	-0.2

Table 8. Centroid Results

Attribute	Cluster_0	Cluster_1	Cluster_2
DKS 1	0.541253645	-1.039692762	0.129568838
DKS 2	0.644168154	-0.620324744	-0.105876369
DKS 3	0.525828083	-0.555727996	-0.065619916
DKS 4	0.442089879	-0.617295287	0.008081471
DSS 1	1.139488503	-1.041422252	-0.210844395
DSS 2	1.174187709	-1.134398456	-0.191443229
DSS 3	0.84715783	-1.393190567	0.104122551
DSS 4	0.93832948	-0.950790205	-0.134334162
MSS	0.798043621	0.059665646	-0.480231056
ESS	-0.242101867	-0.2230281	0.232061669
GR	1.00728978	-0.947683456	-0.174968088
Total	7.815734816	-8.46388818	-0.88948269

3.1.2 *Centroid Visualization of Each Cluster:* The centroids of each cluster show the average values of each attribute used in clustering. Analyzing these centroid values helps to identify the common characteristics of students in each cluster. The chart in Fig. 3 compares the centroid values of Cluster 0, Cluster 1, and Cluster 2 based on attributes. The chart results indicate that Cluster 0 (blue line) has higher centroids for most attributes, signifying that students in this cluster have excellent performance. Cluster 1 (green line) has lower centroids, indicating that students in this cluster have moderate performance. Meanwhile, Cluster 2 (red line) shows more balanced centroid values, suggesting that students in this cluster can be categorized as having good performance.

3.1.3 *K-Means Clustering Results:* The student performance data clustering using the K-Means algorithm produced three clusters representing student performance categories: excellent, good, and moderate. These categories are based on the centroid values of each cluster, as explained in Section 3.1.2. The cluster with the highest centroid value represents the group of students with excellent performance, the cluster with more balanced centroids represents good performance, and the cluster with lower centroid values represents students with moderate performance. Table 9 presents the classification of student performance into three categories: excellent, good, and moderate.

3.1.4 *Number and Ratio of Objects:* The K-Means clustering process applied to student performance data produced three clusters with different numbers of objects and ratios in each cluster. Each cluster represents groups of students who share similar academic performance patterns. Table 10 shows the number of objects in each cluster and their ratio to the overall data. Table 10 reveals that Cluster 2 contains the largest number of objects, with 121 students or 50.21% of the total data, followed by Cluster 0 with 69 students (28.63%) and Cluster 1 with 51 students (21.16%). This distribution helps in understanding the student groupings based on performance patterns resulting from the K-Means analysis.

Table 9. K-Means Clustering Results

Student ID	Performance	Cluster	Average
1	Excellent	cluster_0	93
2	Excellent	cluster_0	88
3	Excellent	cluster_0	90
4	Excellent	cluster_0	95
5	Excellent	cluster_0	95
6	Moderate	cluster_1	86
7	Moderate	cluster_1	86
8	Excellent	cluster_0	92
9	Moderate	cluster_1	81
10	Moderate	cluster_1	84
...	...	...	...
231	Excellent	cluster_0	91
232	Excellent	cluster_0	95
233	Excellent	cluster_0	95
234	Moderate	cluster_1	84
235	Good	cluster_2	87
…	…	...	…
239	Good	cluster_2	85
240	Excellent	cluster_0	93
241	Good	cluster_2	90

Table 10. Number of Cluster Members

Cluster	Number of Students	Ratio (%)
1	69	28,63%
2	51	21,16%
3	121	50,21%

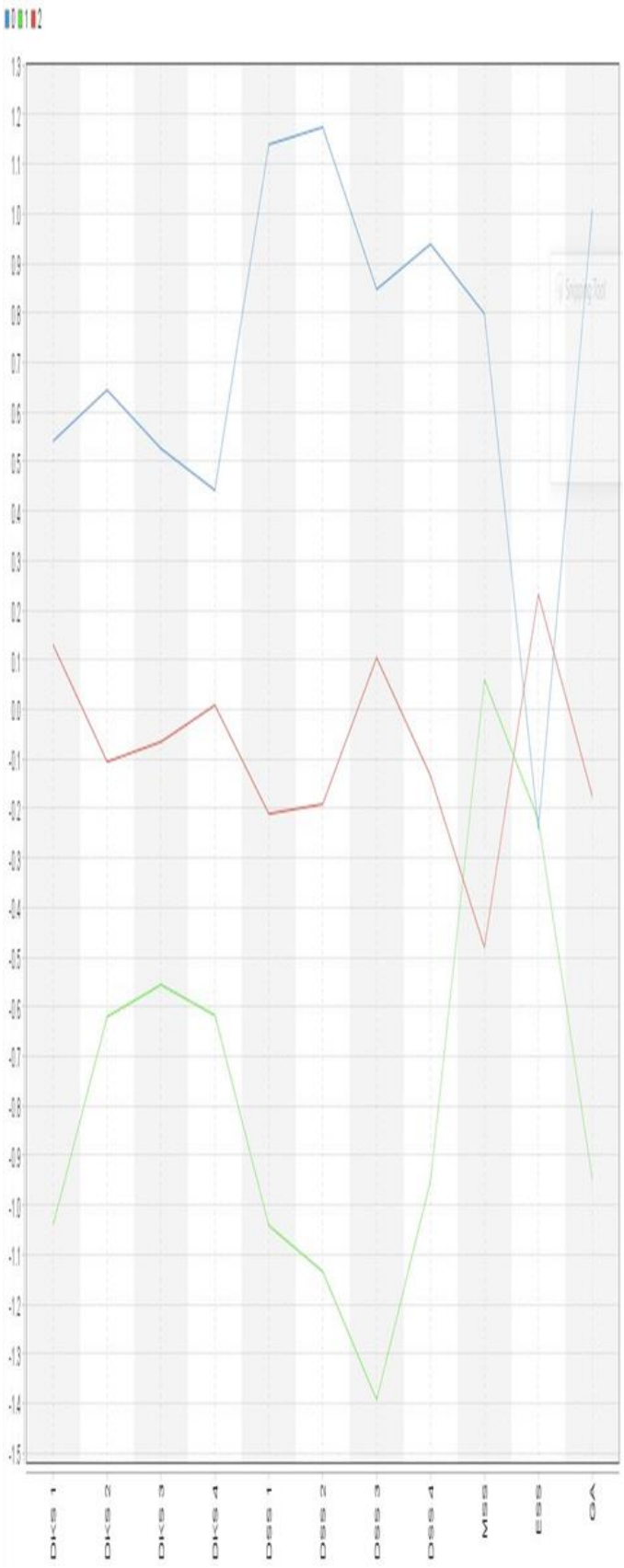

Figure 3. Cluster centroid char

3.1.5 *Bar Chart of Student Distribution:* The distribution of students in each K-Means cluster is visualized through a bar chart, which displays the number of students per cluster. This chart is shown in Fig. 4, generated by the RapidMiner application. The X-axis represents the names of the clusters, while the Y-axis indicates the number of students in each cluster.

3.1.6 *K-Means Clustering Visualization:* The results of the K-Means clustering are visualized using a scatter plot with DKS 1 (Daily Knowledge Score 1) on the X-axis and DKS 2 (Daily Knowledge Score 2) on the Y-axis. The chart in Fig. 5 illustrates the distribution of students based on the formed clusters. In the scatter plot (Fig. 5), each point represents one data object (student), and the color indicates the cluster to which the object belongs. The green color represents Cluster 0, the blue color represents Cluster 1, and the orange color represents Cluster 2. The distribution pattern shows how students are grouped into three main clusters based on DKS 1 and DKS 2. The objects in Cluster 2 (orange) are mostly clustered closely together, indicating that the students in this cluster share similar performance characteristics, and the K-Means process successfully grouped them into a homogeneous category. Although there are some objects slightly outside the main group, their number is insignificant. In contrast, the objects in Cluster 0 (green) and Cluster 1 (blue) are more dispersed, suggesting that students in these clusters have more distinct performance characteristics compared to other clusters.

3.2 Evaluation Results

At the evaluation stage, the Davies-Bouldin Index is used to assess the quality of the clustering produced by the K-Means algorithm by considering the ratio between inter-cluster distance and intra-cluster dispersion. A lower Davies-Bouldin value indicates better clustering performance, with more distinct and compact clusters. This evaluation is conducted for a range of clusters (K), from 2 to 10, using the Loop Parameter operator, which automatically computes the results for each K. Two main operators are employed: K-Means for clustering and Performance (Distance) to calculate the Davies-Bouldin value. The evaluation results are presented in both tables and graphs.

3.2.1 *Davies-Bouldin Index Results:* This section presents the results of the Davies-Bouldin Index calculations for various numbers of clusters (K), ranging from K=2 to K=10. Each Davies-Bouldin Index value reflects the quality of the clustering, with a lower value indicating better clustering results. Table 11 below shows the Davies-Bouldin Index values for each number of clusters.

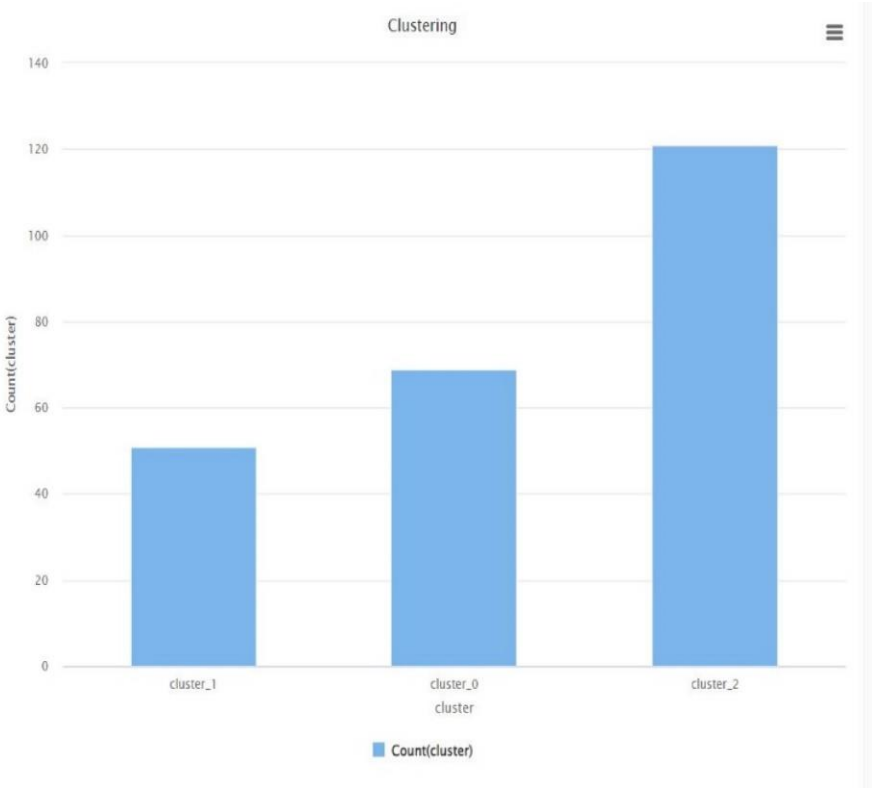

Figure 4. Graph of the number of cluster member

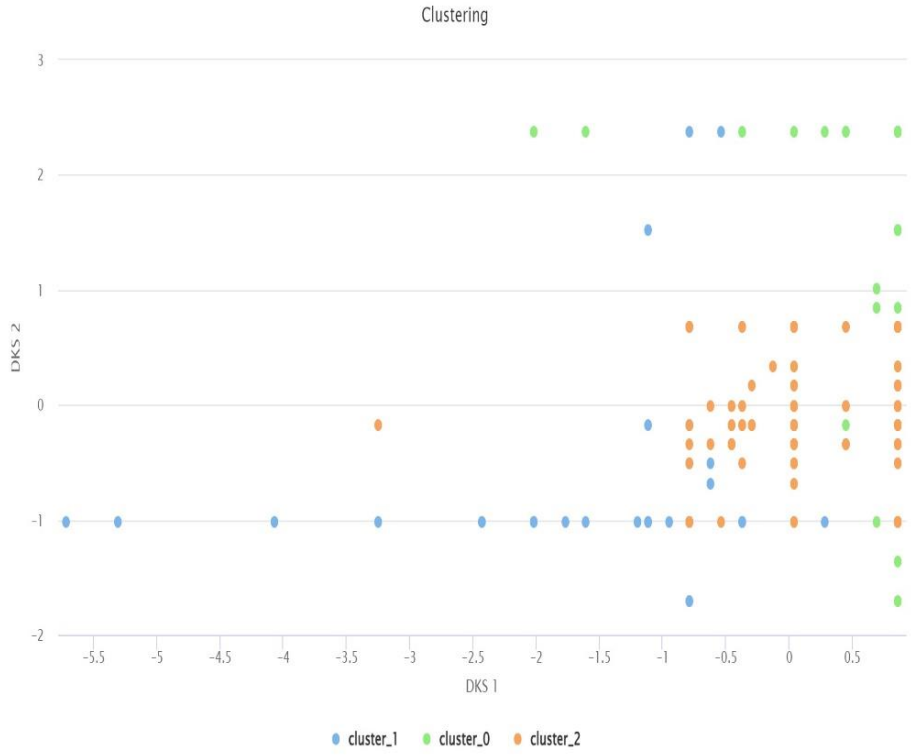

Figure 5. Student distribution chart

Based on the evaluation, the lowest Davies-Bouldin Index value was obtained at K=9 with a value of 1.4786, which technically indicates better clustering than with other numbers of clusters. However, nine clusters were considered too many for this research, which aimed to group students into simpler performance categories. Conversely, K=2 had a Davies-Bouldin Index value of 1.533, but two clusters were deemed insufficient to capture the variation in student performance comprehensively. Therefore, K=3 was selected, even though the Davies-Bouldin Index value was higher (1.713). Selecting three clusters was more appropriate for the analysis's needs to categorize students into three performance levels: excellent, good, and moderate.

3.2.2 *Davies-Bouldin Index Chart:* The section provides a visual representation of the Davies-Bouldin Index calculations for varying numbers of clusters (K) generated by the K-Means algorithm, spanning from K=2 to K=10. Fig. 6 illustrates the changes in Davies-Bouldin Index values as the number of clusters increases, with each point on the chart corresponding to the index value for a specific K. The X-axis represents the number of clusters (K) in the K-Means algorithm, while the Y-axis displays the respective Davies-Bouldin Index values. The chart, presented as a scatter plot, visually demonstrates how clustering quality evolves with increasing K, where lower values suggest improved separation and compactness of the clusters. From the chart, it can be observed how the Davies-Bouldin Index values vary for each K, with K=9 yielding the lowest value. However, as explained in Section 3.2.1, the selection of the number of clusters in this analysis considered other factors besides just the Davies-Bouldin Index value. This visualization provides an overview of how the clustering quality changes as the number of clusters is adjusted.

Table 11. Davies Bouldin Result

The Number of Clusters	Davies Bouldin
2	1.533
3	1.713
4	1.741
5	1.685
6	1.630
7	1.644
8	1.715
9	1.479
10	1.541

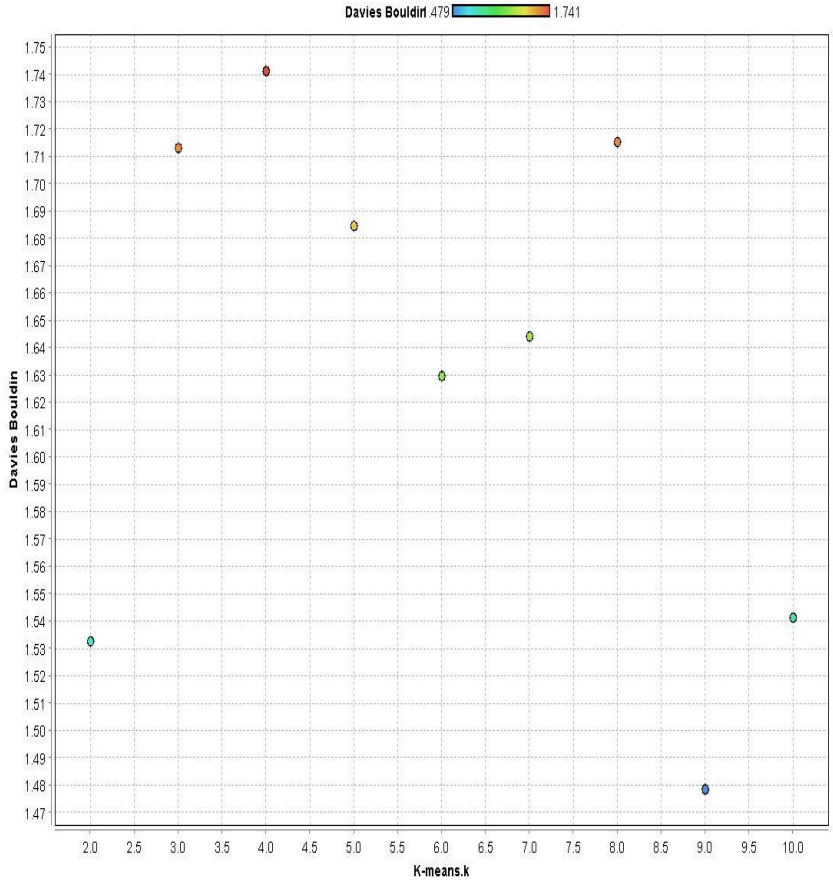

Figure 6. Davies-Bouldin Index Changes

4 CONCLUSION

Based on the graphs and centroid results, students with excellent performance are grouped in cluster 0, with the highest centroid values in most attributes. Cluster 1 consists of students with moderate performance, characterized by lower centroid values, while cluster 2 comprises students with good performance, showing balanced centroid values. K-Means Clustering categorizes students into three academic performance groups: excellent, good, and moderate, reflecting variations in student performance in social studies at Ksatrya Junior High School, Jakarta. A total of 28.63% of students belong to the excellent performance group, 50.21% to the good performance group, and 21.16% to the moderate performance group, with the majority of students in the good performance category.

The evaluation of clustering performance using the Davies-Bouldin Index resulted in a value of 1.713 for K=3, which was chosen as it aligns with the desired three performance categories, despite K=9 having a lower Davies-Bouldin Index of 1.479. The 1.713 value is still considered acceptable in terms of separating the characteristics of each student group.

The results are expected to assist the school in making strategic decisions, such as providing more challenging materials for students with excellent performance, appropriate materials for students with good performance, and repetition-based materials for students with moderate performance. Future developments include optimizing the number of K, applying this method to other subjects, and integrating it with e-learning technology to monitor student progress in real-time.

AUTHOR'S CONTRIBUTION

In the research titled "K-Means Clustering of Social Studies Performance at Ksatrya Junior High School" Tundo provided guidance on the methodology, data analysis methods, and manuscript writing. Syifa Raihanah was responsible for data collection and analysis. Tri Wahyudi contributed suggestions regarding data collection. Sugiyono ensured that the writing adhered to the established methodology. With the guidance and input from the entire team, this research was completed and is expected to make a significant contribution to the field of study.

.

COMPETING INTERESTS

Following the publication ethics of this journal, Tundo, Syifa Raihanah, Tri Wahyudi, and Sugiyono, as the authors

of this journal article, declare that this journal article has no conflict of interest (COI) or competing interests (CI).

ACKNOWLEDGMENT

The author wishes to convey immense thanks to all those who have given important advice, support morally and materially, constructive criticism, and various types of assistance during the research process, which has allowed this research to be completed.

REFERENCES

[1] P. B. N. Simangunsong and M. R. Manalu, "Testing the K-Means Clustering Algorithm in Processing Student Assignment Grades Using the RapidMiner Application," *J. Data Sci.*, vol. 1, no. 2, pp. 51–60, 2023, [Online]. Available: https://ejournal.seaninstitute.or.id/index.php/visualization/article/view/2838/2152

[2] N. O. Malayphone Sonephachanh, "An Approach for Analyzing Student Performance Based on Formative Assessment Scores Using the k-Means Method," *Int. J. Adv. Res. Educ. Soc.*, vol. 6, no. 1, pp. 16–23, 2024, doi: 10.55057/ijares.2024.6.1.2.

[3] U. F. Laili, A. Tanzeh, N. Efendi, and M. Gufron, "K-Means Clustering Method On Academic Advising Management And Early Detection Of Student Dropout (Sequential Explanatory Mixed Method Study At UIN Sunan Ampel Surabaya And IAIN Kediri)," *Int. J. Educ. Res. Soc. Sci.*, vol. 5, no. 2, pp. 335–346, 2024, doi: 10.51601/ijersc.v5i2.744.

[4] I. W. P. 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," *Enrich. J. Multidiscip. Res. Dev.*, vol. 1, no. 7, pp. 397–411, 2023, doi: 10.55324/enrichment.v1i7.64.

[5] A. F. Arwani, "Intelligent Learning Achievement Prediction System Using K-Means Algorithm at UPT Education Unit SMPN 5 BLITAR," *J. Students Acad. Res.*, vol. 9, no. 1, pp. 215–222, 2024, doi: https://doi.org/10.35457/josar.v9i1.3100.

[6] M. A. Jabbar, Fi. E. Silmi, and A. T. Satria, "Analysis of Teaching and Learning Activities to K-Means Clustering Method (Case Study of SMK Maarif Al-Mizan)," *J. Multimed. Technol. Appl. Softw.*, vol. 1, no. 1, pp. 28–35, 2024, [Online]. Available: https://ejournal.poltekmi.ac.id/index.php/jmtas/article/view/11/6

[7] C. Yu and Y. Wang, "College Student Management System Based on K-means Clustering Algorithm," *Int. J. New Dev. Educ.*, vol. 4, no. 2, pp. 28–33, 2022, doi: 10.25236/ijnde.2022.040206.

[8] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[9] Kertanah, W. P. Nurmayanti, S. R. Aini, L. M. Amrullah, and M. Sya'roni, "Comparison of Algorithms K-Means and DBSCAN for Clustering Student Cognitive Learning Outcomes in Physics Subject," *Kappa J. Phys. Phys. Educ.*, vol. 7, no. 1, pp. 251–255, 2023, doi: 10.29408/kpj.v7i2.18428.

[10] E. L. Cahapin, B. A. Malabag, C. S. S. Jr, J. L. Reyes, G. S. Legaspi, and K. L. Adrales, "Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms," *Bull. Electr. Eng. Informatics*, vol. 12, no. 6, pp. 3647–3656, 2023, doi: 10.11591/eei.v12i6.4849.

[11] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electron.*, vol. 9, no. 8, pp. 1–12, 2020, doi: 10.3390/electronics9081295.

[12] F. A. Syah, S. P. Hasugian, M. V. A. Adafmi, and A. D. Sibarani, "The Implementation of the K-Means Clustering Algorithm for Awarding Scholarships to Outstanding Students," *J. Comput. Intell. Informatics*, vol. 1, no. 1, pp. 9–16, 2024, [Online]. Available: https://journal.unilak.ac.id/index.php/ComniTech/article/view/21127/6503

[13] R. F. Anggraini and S. Sau'da, "Stream Clustering for Selection Recommendations Using K-Means Algorithm: A Case Study in the Informatics Study Program," *J. Inf. Syst. Informatics*, vol. 5, no. 4, pp. 1274–1287, 2023, doi: 10.51519/journalisi.v5i4.576.

[14] A. P. Adistya, N. Lutfiyani, P. Tara, Rifaldi, R. Adriyan, and P. Rosyani, "Klasterisasi Menggunakan Algoritma K-Means Clustering Untuk Memprediksi Kelulusan Mata Kuliah Mahasiswa," *OKTAL J. Ilmu Komput. dan Sci.*, vol. 2, no. 8, pp. 2301–2306, 2023, [Online]. Available: https://journal.mediapublikasi.id/index.php/oktal/article/view/1610/1857

[15] P. Rosyani and F. Syawali, "Application of Advanced Class Determination System Using K-Means Clustering Method (Case Study: SMK Al-Badar Balaraja)," *Int. J. Integr. Sci.*, vol. 2, no. 10, pp. 1557–1570, 2023, doi: 10.55927/ijis.v2i10.6347.

[16] A. F. M. Nafuri, N. S. Sani, N. F. A. Zainudin, A. hadi abd Rahman, and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education," *Appl. Sci.*, vol. 12, no. 19, 2022, doi: 10.3390/app12199467.

[17] B. Sutara, F. Vulture, and R. Novianti, "Application of K-Means algorithm with CRISP-DM method in student data analysis as a support for promotion strategy," *SIDE Sci. Dev. J.*, vol. 1, no. 1, pp. 1–7, 2024, [Online]. Available: https://ojs.arbain.co.id/index.php/side/article/view/6/6

[18] M. Wati, W. H. Rahmah, N. Novirasari, Haviluddin, E. Budiman, and Islamiyah, "Analysis K-Means Clustering to Predicting Student Graduation," *J. Phys. Conf. Ser.*, vol. 1844, no. 1, pp. 1–8, 2021, doi: 10.1088/1742-6596/1844/1/012028.

[19] T. Wahyudi and T. Silfia, "Implementation of Data Mining Using K-Means Clustering Method To Determine Sales Strategy in S&R Baby Store," *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 1, pp. 93–103, 2022, doi: 10.37385/jaets.v4i1.913.

[20] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, "K-means algorithm for clustering of learners performance levels using machine learning techniques," *Int. Inf. Eng. Technol. Assoc.*, vol. 35, no. 1, pp. 99–104, 2021, doi: 10.18280/ria.350112.

[21] S. N. Wahyuni, N. N. Khanom, and Y. Astuti, "K-Means Algorithm Analysis for Election Cluster Prediction," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 1–6, 2023, doi: 10.30630/joiv.7.1.1107.

[22] M. A. Saputra and S. Harini, "Java Island Health Profile Clustering using K-Means Data Mining," *Int. J. Inf. Commun. Technol.*, vol. 8, no. 1, pp. 1–9, 2022, doi: 10.21108/ijoict.v8i1.606.

[23] M. F. J. Muttaqin, "Cluster Analysis Using K-Means Method to Classify Sumatera Regency and City Based on Human Development Index Indicator," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 967–976, 2022, doi: 10.34123/semnasoffstat.v2022i1.1299.

[24] S. Kriˇzanic, "Educational data mining using cluster analysis and decision tree technique : A case study," *Int. J. Eng. Bus. Manag.*, vol. 12, pp. 1–9, 2020, doi: 10.1177/1847979020908675.

[25] M. Al Ghifari and W. T. H. Putri, "Clustering Courses Based On Student Grades Using K-Means Algorithm With Elbow Method For Centroid Determination," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 8, no. 1, pp. 42–46, 2023, doi: 10.25139/inform.v8i1.4519.

[26] H. Hasanah, N. A. Sudibyo, and R. M. Galih, "Data Mining Using K-Means Clustering Algorithm for Grouping Countries of Origin of Foreign Tourist," *Basic Appl. Sci. Conf.*, vol. 2021, pp. 88–94, 2021, doi: 10.11594/ nstp.2021.1112.

[27] L. Restuono, A. P. U. Siahaan, R. F. Wijaya, Z. Sitorus, and M. Iqbal, "International Journal of Computer Sciences and Mathematics Engineering Analysis and Exploration of Clustering Algorithms for New Student Segmentation," *Int. J. Comput. Sci. Math. Eng.*, vol. 3, no. 1, 2024, doi: https://doi.org/10.61306/ijecom.v3i1.61.

[28] A. J. E. Sakalessy and H. D. Purnomo, "Assessing Employee Performance in the Information Technology Department Using K-Means Clustering: A Case Study Approach," *J. Inf. Syst. Informatics*, vol. 6, no. 1, pp. 170–186, 2024, doi: 10.51519/journalisi.v6i1.653.

[29] O. W. B. Jnr, "K- Means, Clustering Algorithm for Student's Selection and Performance Prediction," *Int. J. Sci. Technol. Res.*, vol. 10, no. 10, pp. 35–39, 2021, [Online]. Available: https://www.ijstr.org/final-print/oct2021/K-means-Clustering-Algorithm-For-Students-Selection-And-Performance-Prediction.pdf

[30] Zuliani, R. Buaton, and S. Ramadani, "Student Character Grouping Based on Six Dimensions of Pancasila Student Profile Using Clustering Method (Case Study of SMK Swasta Setia Budi Binjai)," *Int. J. Informatics, Econ. Manag. Sci.*, vol. 2, no. 2, pp. 130–140, 2023, doi: 10.52362/ijiems.v2i2.1202.