Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 2 Issue 1 June, 2024, pp. 21-28
https://idjams.com/index.php/jms/index

Application of Cluster Analysis and Correlation between Mathematics and Natural Sciences Subject Based on Student Test Scores Using K-Means Clustering

Sekar Pangesti[1,*], Nurul Farisah Binti Roslan[2], Andreansyah[3]

[1] *Universitas Singaperbangsa Karawang, Jl. HS. Ronggowaluyo, Karawang, 41361, Indonesia*

[2] *Universiti Teknologi MARA, Cawangan Kelantan, Kota Bharu, 15050, Malaysia*

[3] *Jiangsu Agri Animal Husbandry Vocational Collage, 8 East Phoenix Road, Taizhou, Jiangsu, 225300, China*

[*]*Corresponding author: 2010631050104@student.unsika.ac.id*

Copyright: ©2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License-(CC-BY-SA) (https://creativecommons.org/licenses/by-sa/4.0/)

Article Info

Article history:
Received: June 29, 2024
Revised: September 24, 2024
Accepted: October 5, 2024
Available online: June 1, 2025

DOI:
https://doi.org/10.71385/idjams.v2i1.8

Abstract

Educational unit examinations are a means of supporting decisions to determine students' abilities in mastering all the material that has been taught. education unit examinations also have an impact on students' ability to continue to the next level of education, besides that they can also improve the quality of school education. Based on the background of the problem, this researcher aims to group learning outcomes for mathematics and natural sciences subjects based on educational unit exam scores using the k-means clustering algorithm, in addition to finding out the relationship between mathematics and natural sciences subject learning outcomes using Spearman rank correlation analysis. The population in this study was all 39 students in class IX of SMP Plus Berkualitas Lengkong Mandiri for the 2023/2024 academic year. The sample in this study used saturated sampling. The research results show that through clustering using the k-means clustering algorithm, it was found that cluster 1 with the high category had 21 students, cluster 2 with the medium category had 12 students and cluster 3 with the low category had 6 students. Based on the Spearman rank correlation analysis of the entire sample, a low correlation was obtained between mathematics results and natural sciences results based on educational unit exam scores of 0.608 and a coefficient of determination of 36.9%. Meanwhile, the Spearman rank correlation test on all cluster 1 samples showed a moderate relationship between mathematics learning outcomes and natural sciences subject learning outcomes of 49.5%.

Keywords: Cluster Analysis; Student Test; Natural Science; Mathematics

I. Introduction

Mathematics is a subject that is no longer foreign to the world of education, where mathematics is a mandatory subject taught from elementary school to higher education [1]. This is because mathematics has an important role not only in the world of education, but also in life [2]. According to Nuraeni, et al [3] one of the importance of mathematics in life is that it is useful in solving and resolving problems that are often encountered in everyday life. This is in line with Kline [4] that mathematics is not knowledge that can stand alone, but mathematics can help humans understand and overcome social, economic and natural problems. The importance of mathematics in life requires students to master mathematics well [5].

Although previous studies have explained the importance of mathematics in life, there are still things that need to be studied, especially related to understanding the concept of mathematics with

natural sciences. With this understanding of the concept, it can contribute to a deeper understanding of how mathematics can solve problems in the field of natural sciences.

Mathematics and natural sciences are considered difficult to understand by some students. Mathematics is difficult for students because mathematics uses many formulas [6]. Meanwhile, natural sciences are considered difficult by students because natural sciences have a lot of memorization apart from formulas [7]. Mathematics and science are two sciences that have roles and relationships that are closely related to each other, both in language and calculations. Mathematics helps science in carrying out calculations, especially in the branches of physics, chemistry and biology [8]. Meanwhile, science helps mathematics, such as explaining units and their conversions.

Another connection between mathematics and natural sciences is that there are theories or formulas that were born in science after mathematics discovered them first, for example: Geometry obtained by C.F.B. Reimann in 1854 and the inverian theory by mathematician A. Cayley gave birth to Albert Einstein's theory of gravity, a mathematical theory of boundary value problems discovered by J.C.F. Stern and J. Lionville in 1830, gave birth to the theory of wave mechanics, and a revolution in modern physics that began with the work of W. Heisenberg and P.A.M. Dirac in 1925 because of the mathematical matrices discovered by Cayley in 1858 [8].

In modern times like today, it is impossible to ignore the role of mathematics in making machines, bridges, dams, and even space travel, including the role of mathematics, including natural sciences which will always require mathematics in their development. Thus, mathematics in natural sciences (IPA) has a role and close relationship both in terms of language and calculations and so on.

Based on these problems, researchers are interested in conducting research regarding the relationship between mathematics ability and natural sciences ability based on the results of educational unit exams. If there is a relationship, what does the connection look like, is it a unidirectional relationship or a unidirectional relationship? Apart from that, researchers are also interested in grouping according to students' abilities using k-means clustering. According to Yudhanegara et al. [9], k-means clustering is a non-hierarchical data clustering method that groups data in the form of one or more clusters. Data that has the same characteristics is grouped in one cluster and data that has different characteristics is grouped in another cluster so that the data in one cluster has a small level of variation. Cluster analysis using the k-means method is starting to be widely used in the field of education because it is considered an accurate and efficient method for grouping various learning approaches and student abilities.

II. Method

The research technique used is k-means clustering and correlational algorithm grouping analysis that focuses on the relationship between the two variables, so the method used in this study is Spearman rank correlation analysis. In this study, the approach used is quantitative in the form of data set processing. The data used in the study are the results of the Education Unit Examination (USP) for mathematics and natural sciences subject for grade IX from private schools in South Tangerang, so the data sources used are secondary data.

Population as the entire research object, in this study the population was all class IX students of SMP Plus Berkualitas Lengkong Mandiri. It is a Junior High School. According to Sugiyono [10], the sample is part of the number and characteristics of the population. The sample in this study was determined using saturated sampling, namely determining the sample if all members of the population were used as samples [11]. The number of samples taken in this study was 39 students.

The type of instrument used in this research is collecting data sets directly through the school, because it is in accordance with the approach used and the objectives of this research. The data obtained was then clustered using the k-means clustering algorithm which was divided into three clusters, namely high, medium and low. The steps for clustering using the k-means method are as follows:

1. Determine the k value as the number of clusters you want to form.
2. Initialize k as a centroid that can be generated randomly.

Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 2 Issue 1 June, 2024, pp. 21-28
https://idjams.com/index.php/jms/index

3. Calculate the distance of each data to each centroid using the Euclidean Distance equation as follows:
4. $$d(x_i, x_j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |(x_{ip} - x_{jp}|^2)} \quad (1)$$

   x_i, x_j are two data whose distance is calculated and p is the dimension of the data used. Determining the cluster center point can be seen from the following equation:

$$C_{m(q)} = \frac{1}{n_m}\sum_{i=1}^{n_m} x_{i(q)} \quad (2)$$

   where
   $C_{m(q)}$: center of the mth group of variable p
   m : 1, 2, ..., k,
   n_m : number of objects in the mth group,
   k : number of clusters,
   q : 1, 2, ..., p,
   x_i : observation value of the i-th object of the q-th variable,
   i : 1, 2, ..., n_m.
5. Group each data based on the closest distance between the data and their respective centroids
6. Determine the position of the new centroid (k)
7. Return to step 3 if the position of the new centroid with the old centroid changes and stop if the new centroid with the centroid does not change.

After clustering, the data will be tested again using the Spearman rank correlation test of all samples, cluster 1 samples, cluster 2 samples and cluster 3 samples to determine the relationship between mathematics learning outcomes and Indonesian language learning outcomes based on the scores obtained from the education unit exams. In determining the Spearman rank correlation coefficient value, use the following formula:

$$r_s = \frac{\sum x^2 + \sum y^2 - \sum {d_1}^2}{2\sqrt{\sum x^2 \sum y^2}} \quad (3)$$

where $\sum x^2 = \frac{n^2-n}{12} - \sum T_x$, $\sum y^2 = \frac{n^2-n}{12} - \sum T_y$, $\sum T_x = \sum T_y = \frac{t^2-t}{12}$, n = the number of data, and t = the same observation.

Testing the significance level of Spearman rank correlation uses the following hypothesis:
H_0: there is no significant relationship between mathematics learning outcomes and natural sciences learning outcomes.
H_a: there is a significant relationship between mathematics learning outcomes and natural sciences learning outcomes.
Then, the correlation coefficient value can be interpreted as the level of correlation that exists to determine the strength of the relationship between mathematics learning outcomes and natural science learning outcomes. Based on Guiford's criteria [11], the level of direction of the relationship between mathematics learning outcomes and natural science learning outcomes is used, as follows:

Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 2 Issue 1 June, 2024, pp. 21-28
https://idjams.com/index.php/jms/index

Table 1. Interpretation of Correlation

Correlation Coefficient Value	Interpretation
$0{,}00 < r_s < 0{,}20$	Very weak relationship (negligible)
$0{,}20 \leq r_s < 0{,}40$	Low correlation
$0{,}40 \leq r_s < 0{,}70$	Moderate/fair relationship
$0{,}70 \leq r_s < 0{,}90$	Strong/high relationship
$0{,}90 \leq r_s \leq 1{,}00$	Very strong/high relationship

Next, to find out the direction of the correlation, you can look at the correlation coefficient number. The correlation coefficient value is between +1 to -1. If the correlation coefficient value is positive then the relationship between the two variables is in the same direction, meaning that if the value of one variable increases then the other variable also increases and vice versa if the value of one variable decreases then the other variable will also decrease. Then, if the correlation coefficient value is negative then the relationship between the two variables is not in the same direction, which means that if the value of one variable increases then the value of the other variable decreases and vice versa if the value of one variable decreases then the value of the other variable increases. If the correlation coefficient value is 0 then the relationship between the two variables has no direction.

The test statistics in this research use the t-test with a significance level of 5%, for the test statistics use the following formula:

$$t_{statistics} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s{}^2}} \tag{4}$$

with test criteria if $t_{statistics} \leq t_{table}$, then H_o is not rejected and H_a is rejected,if $t_{statistics} > t_{table}$, then H_o is rejected and H_a is not rejected.

III. Results And Discussion

In order to achieve the objectives of this research, the researcher will carry out testing using educational unit exam data for mathematics subjects which consists of 30 multiple choice questions, and 5 essay questions, and science consisting of 40 multiple choice questions and 5 essay questions, at one of the junior high schools in South Tangerang, namely SMP Plus Berkualitas Lengkong Mandiri for the academic year 2023/2024 by carrying out clustering first.

At this stage, the scoring results are grouped, the data will be divided into 3 clusters, namely, cluster 1 in the high category, cluster 2 in the medium category and cluster 3 in the low category. The following are the clustering stages using the k-means clustering method:

1) Determining the number of clusters, the researcher chooses 3 clusters (k=3) to be created. the number of clusters must be smaller than the number of data (k < n)
2) Determine the centroid of each cluster, to determine the initial centroid (initial centroid) is done randomly. the following centroid data is obtained:

Table 2. Initial Centroid

Centroid	Mathematics	Natural Sciences
C1	74	60
C2	76	100
C3	86	79

3) Calculation of the cluster center distance, to measure the distance between data and the centroid, use the Euclidean distance in Equation 1.
   The data distance to centroid 1 is:
   $d(1,1) = \sqrt{|74-74|^2 + |60-60|^2} = 0.00$

Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 2 Issue 1 June, 2024, pp. 21-28
https://idjams.com/index.php/jms/index

d(2,1) = $\sqrt{|80-74|^2 + |60-60|^2}$ = 6.00

d(3,1) = $\sqrt{|74-74|^2 + |68-60|^2}$ = 8.00

And so on until the 39th data, then calculate the distance between the data and centroid 2 as follows:

d(1,2) = $\sqrt{|74-76|^2 + |60-100|^2}$ = 40.05

d(2,2) = $\sqrt{|80-76|^2 + |60-100|^2}$ = 40.20

d(3,2) = $\sqrt{|74-76|^2 + |68-100|^2}$ = 32.06

And so on until the 39th data, then calculate the distance between the data and centroid 3 as follows:

d(1,3) = $\sqrt{|74-86| + |60-79|^2}$ = 22.47

d(2,3) = $\sqrt{|80-86| + |60-79|^2}$ = 19.92

d(3,3) = $\sqrt{|74-86| + |68-79|^2}$ = 16.28

and so on until data 39.

4) Distance calculation results and clustering results

To group data based on the closest distance to the centroid, it can be seen from the minimum distance for each data. Then determine a new centroid to carry out the 2nd iteration process with the same steps as the 1st iteration process. Then look at the cluster results in the 1st iteration and 2nd iteration whether there are changes or not. If there is no change then the iteration process is stopped, but if there is a change then it continues with the next iteration process until the cluster results have no change. The following are changes in cluster centers obtained from the data using IBM SPSS Statistics 29 software.

Table 3. Change in Cluster Centers

Iteration to	Cluster 1	Cluster 2	Cluster 3
1	3.048	6.333	3.371
2	0.000	0.000	0.000

If you look at the change in the cluster center of the 2nd iteration, the value is 0.000, meaning there is no change in any data in each cluster from the 1st and 2nd iterations, then the iteration process is stopped. With the final cluster center as follows:

Table 4. Final Cluster Centers

Variable	Cluster 1	Cluster 2	Cluster 3
Mathematics	77	82	83
Natural Sciences	61	100	80

The following are the results of grouping using the k-means clustering method.

Table 5. Grouping Results of the K-Means Clustering Method

Cluster	Number of Students	Category
1	21 Students	High
2	12 Students	Medium
3	6 Students	Low

Based on Table 5, it is found that cluster 1 has 21 students in the high category, cluster 2 has 12 students in the medium category and cluster 3 has 6 students in the low category.

Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 2 Issue 1 June, 2024, pp. 21-28
https://idjams.com/index.php/jms/index

To determine the existence of a significant and positive relationship between mathematics learning outcomes and natural sciences learning outcomes throughout the data, the Spearman rank correlation calculation results were obtained using IBM SPSS statistics software as follows.

Table 6. Spearman Rank Correlation Calculation Results for All Samples

			Mathematics	**Natural Sciences**
Spearman' rho	**Mathematics**	Correlation Coefficient	1.000	.608**
		Sig. (2-tailed)		.000
		N	39	39
	Natural Sciences	Correlation Coefficient	.608**	1.000
		Sig. (2-tailed)	.000	
		N	39	39

Based on Table 6, it can be seen that the correlation between mathematics and natural sciences variables in the entire sample shows a positive number of 0.608, which means a unidirectional relationship. This means that if the mathematics education unit exam results increase, then science also increases, and vice versa. Based on Table 1. The correlation coefficient value obtained has a moderate/fair relationship. Next, to test the significance of the correlation coefficient value, use the t test as follows:

$$t_{statistics} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s{}^2}} = \frac{0.608\sqrt{39-2}}{\sqrt{1-(0.608)^2}} = 4.658$$

From these calculations, the $t_{statistics}$ value is 4.658, then compare it with the t_{table} value using the condition α = 5% with a sample 39 then degree of freedom = 39-2 = 37 so the t_{table} value is 1,687. From these calculations it is found that the value of $t_{statistics} > t_{table}$, then H_0 is rejected. Thus, there is a significant correlation between the results of the mathematics education unit exam and natural sciences. The coefficient of determination value shows the magnitude of the influence of the mathematics education unit exam results on natural sciences on all data in the sample, namely sebesar $r_s{}^2$ = 0.369 or 36.9% and vice versa, while the other 63.1% is influenced by other factors.

To find out whether there is a significant and positive relationship between the results of the mathematics and science education unit exams in cluster 1, the results of the Spearman rank correlation calculation using IBM SPSS statistics software are obtained as follows:

Table 7. Results of Cluster 1 Spearman Rank Correlation Calculation

			Mathematics	**Natural Sciences**
Spearman' rho	**Mathematics**	Correlation Coefficient	1.000	-.399
		Sig. (2-tailed)		.074
		N	21	21
	Natural Sciences	Correlation Coefficient	-.399	1.000
		Sig. (2-tailed)	.074	
		N	21	21

Based on Table 6, it can be seen that the correlation between mathematics and natural sciences variables in the entire sample shows a positive number of 0.608, which means a unidirectional relationship. This means that if the mathematics education unit exam results increase, then science also

increases, and vice versa. Based on Table 1. The correlation coefficient value obtained has a moderate/fair relationship. Next, to test the significance of the correlation coefficient value, use the T test as follows:

$$t_{statistics} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s{}^2}} = \frac{0.704\sqrt{39-2}}{\sqrt{1-(0.704)^2}} = 6.029$$

From these calculations, the $t_{statistics}$ value is 6,029, then compare it with the t_{table} value using the condition α = 5% with a sample of 21 then degree of freedom = 21-2 = 17 so the t_{table} results is 1.729. From these calculations it is found that the value of $t_{statistics} > t_{table}$, then H_0 is rejected. Thus, there is a significant correlation between the results of the mathematics education unit exam and natural sciences. The coefficient of determination value shows the magnitude of the influence of the mathematics education unit exam results on science on all data in the sample, namely $r_s{}^2$ = 0.495 or 49,5% and vice versa, while the other 51.5% is influenced by other factors.

IV. Conclusion

Based on the results of data processing of the grouping of the results of the mathematics and natural sciences education unit exams, it was obtained that cluster 1 was in the high category with 21 students, cluster 2 was in the medium category with 12 students and cluster 3 was in the low category with 6 students. In general, the results of all samples showed a significant correlation between mathematics results and natural sciences results. The magnitude of the influence of the mathematics education unit exam results on the science education unit exam results of all samples was 36.9% and vice versa, while the remaining 63.1% was influenced by other factors. Meanwhile, specifically the results of all samples of cluster 1 (high category) showed a significant correlation between mathematics results and natural sciences results. Therefore, the results of this study can be used as a reference for other similar studies. Furthermore, to determine the optimal number of clusters, it is necessary to use the Elbow method and the variables analyzed can be used in all subjects in the school in order to better understand students' abilities in each subject.

References

[1] N. Amalia and E. Unaenah, "Analysis of Mathematics Learning Difficulties in Grade 3 Elementary School Students," *Attadib: J. Elem. Educ.*, vol. 2, no. 2, pp. 123–133, 2018, DOI: 10.32507/attadib.v2i2.414

[2] S. Natalia, "Realistic Mathematics Education: A Step to Educate Mathematical Thinking," *J. Dinamika Pendidik.*, vol. 10, no. 1, pp. 267–282, 2017, DOI: 10.51212/jdp.v10i1.413

[3] R. Nuraeni, S. G. Ardiansyah, and L. S. Zanthy, "Social Arithmetic Mathematics Problems in Story Form: How to Describe Students' Errors in Answers?" *Teorema: Teori dan Riset Matematika*, vol. 5, no. 1, pp. 61–68, 2020, DOI: 10.25157/teorema.v5i1.3345

[4] S. H. Noer, "*Mathematics Learning Strategies."* Matematika, 2017.

[5] D. Kurniawati and A. Ekayanti, "The Importance of Critical Thinking in Mathematics Learning," *PeTeKa*, vol. 3, no. 2, pp. 107–114, 2020.

[6] R. A. Andriansyah, A. N. Mua'arif, D. Nataliasari, S. Rahmin, S. Kurniawati, and D. Darmadi, "Difficulties in Online Mathematics Learning during the COVID-19 Pandemic Among Class VIII Junior High School Students," *J. Educ. Couns.*, vol. 3, no. 2, pp. 67–71, 2021, DOI: 10.31004/jpdk.v3i2.1787

[7] O. K. Bella and B. S. Bachri, "The Influence of the Problem Based Learning Model on Learning Outcomes in Material Density in Class VII Natural Science Subjects in Junior High Schools in Surabaya," *J. Educ. Technol. Students*, vol. 10, no. 10, 2020.

[8] D. I. P. Alam and S. Silaban, "*Basics of Mathematics Education."* Smart Hope Publisher, 2017.

[9] M.R. Yudhanegara, S.W. Indratno, and R.K.N. Sari, "Clustering for Item Delivery Using Rule-K-Means, *J. Indones. Math. Soc.*, vol. 26, no. 2, pp. 185–191, 2020, DOI: 10.22342/jims.26.2.871.185-191.

[10] Sugiyono, "*Quantitative, Qualitative and R&D Research Methods*." Alphabeta, 2018.

[11] K. E. Lestari and M. R. Yudhanegara, "*Mathematics Education Research*." *PT Refika Aditama*, 2015.