Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. Sec. 5. No. 1; e-ISSN: 2808-4519

Clustering of Extracurricular Interest at SMP Negeri 5 Kota Binjai

Adhe Ananda Virginia[1]*, Novriyenni[2], Ratih Puspadini[3]

[1,2,3]STMIK Kaputama,Indonesia
virginiaadheananda@gmail.com1*, novriyenni.sikumbang@gmail.com2 , puspadini.ratih@gmail.com3

Abstract

This study grouped the interest of SMP Negeri 5 Binjai City students in extracurricular activities using the clustering method based on 2018–2024 data with variables of activity type, activeness, and achievement. The goal is to identify patterns of interest and utilize them for program management and development. The results are expected to help schools develop effective coaching strategies according to the characteristics of students. Education shapes character and develops students' potential not only academically, but also through extracurriculars. At SMP Negeri 5 Binjai City, extracurricular participation is not evenly distributed, affecting the effectiveness of supervisors and the development of activities. This study uses the clustering method to group students based on interests, activeness, and achievements, thereby helping schools manage and develop extracurriculars more effectively. This research is carried out in a structured manner through several stages: identification of problems to determine the focus of the research, collection of supporting and main data, study of related theories, analysis of data according to variables, testing and implementation of results, and evaluation to conclude findings and provide suggestions. This framework ensures that research is directed to produce useful results. This study succeeded in grouping the extracurricular interests of SMP Negeri 5 Binjai City students using the k-means algorithm with variables of activity type, activeness, and achievement through three cluster scenarios. Three clusters are sufficient for general strategies, while four or five clusters provide more specific coaching details, helping schools organize student motivational strategies, facilities, mentors, and programs.

Keywords: *Clustering, K-means, Student interest, Extracurricular, Achievement*

1. Introduction

Education is one of the most important things in building human resources who are cultured, have integrity, responsibility and are ready to face challenges in life because they form good character. In the world of education, students do not only focus on the academic aspect, but also include the development of students' talents, interests, and social skills. One of the talent development forums can be through extracurricular [1].

The absence of a database system that assists schools in understanding student preferences can worsen the situation and lead to an imbalance of interest in extracurriculars. In overcoming this problem, there needs to be a method as a solution, namely by using data mining. Data mining aims to determine patterns, relationships, or information that may not be apparent from the data itself, leading to deeper understanding [2].

Another research that applies the clustering method is conducted by a journal entitled "Implementation of Data Mining for Clustering of Inpatient Data with K-Means Clustering Algorithm". The study succeeded in grouping patient data based on age and disease diagnosis to support the improvement of health services.[3].

2. Literature Review

2.1. Interest

Interest is an inner drive that encourages them to engage in an activity actively and enjoyably, because individuals feel attracted to the activity. In the context of learning, learning interests reflect students' interest in learning activities that directly affect the achievement of learning outcomes. When students have a high interest, their involvement in learning tends to increase, as well as the results obtained [4].

2.2. Extracurricular

Extracurricular activities are extracurricular activities that are guided by schools to develop students' talents, interests, and skills. Based on the Regulation of the Minister of Education and Culture Number 62 of 2014, this activity aims to support the academic, personality, and social development of students. Therefore, it can be concluded that extracurricular activities are an important part of the student learning

experience outside of the formal curriculum. This activity provides opportunities for students to gain meaningful experiences outside of the classroom, form their personalities, and develop their potential. In this diverse setting, students not only learn academically, but also grow in social, emotional, and other life skills [5].

2.3. Matlab

Matlab is a software specifically developed to support programming, analysis, and calculation activities of a technical and mathematical nature, especially related to matrix shapes. The name Matlab itself comes from the abbreviation Matrix Laboratory because initially this device was designed to solve linear algebra problems in the form of matrices. The first version has been introduced since 1970 and until now Matlab continues to undergo improvements both in terms of features and computing performance [6] .

2.4. Data Mining

Data mining is the process of extracting valuable information from large data sets using various analysis techniques, such as machine learning, statistics and artificial intelligence. This process aims to find hidden patterns in the data that can be used for decision-making in various fields [7].

2.5. Clustering

Clustering is a way of grouping data into several groups or clusters based on what they have in common. This process separates data or vectors into groups according to their characteristics. Data that has similar characteristics will enter the same cluster, while data with different characteristics will enter a different cluster. The main goal is to create data groups containing similar objects as much as possible. Typically, these data are represented as points in multidimensional space [8] .

3. Research Methods

Data mining is divided into integral parts of Knowledge Discovery in Database (KDD) which is a step in the process of finding patterns contained in each information. In general, the KDD process consists of several stages, namely:

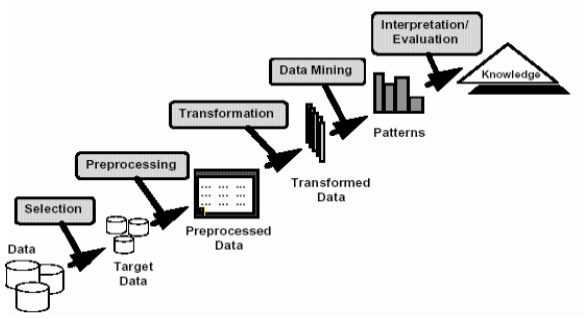

Fig. 1: KDD process

1. Data Selection: To select relevant data from a large operational database.
2. Data Cleansing (Pre-processing/Cleaning): Cleaning of the data to be analyzed.
3. Data Transformation: To format the data in accordance with the needs of data mining analysis.
4. Data Mining: Looking for patterns or useful information from the data that has been prepared.
5. Interpretation and Evaluation: Interpret the results of data mining in a form that is easy for the end user to understand.

3.1. Data Source

The source of the research data on clustering extracurricular interests of SMP Negeri 5 Binjai City students was obtained from the 2018-2024 student extracurricular data archive collected directly from the school. Data access is obtained with official permission from the school authority and the data format is adjusted to meet the requirements of the k-means algorithm.

3.2. Population and Sample

The identified population consists of all students at SMP Negeri 5 Binjai City who have a record of interest or involvement in extracurricular activities in 2018-2024. This p-research sample uses the total sampling method, namely all extracurricular interest data of SMP Negeri 5 Binjai City students for the 2018-2024 period which are recorded in full in the school archives.

3.3. Data Selection

Data collection was carried out by archiving the extracurricular activities of SMP Negeri 5 Binjai City students. The data was obtained directly from the school's official records, which include students involved in extracurricular activities in 2018-2024. The data collection process is structured to ensure that all the data obtained is relevant and complete for clustering analysis. The data was then processed to apply a k-means algorithm that aims to classify students' extracurricular interests based on student activity and achievement.

3.4. Data Analysis

Data analysis was carried out using the k-means clustering algorithm. K-means clustering is one of the most popular and simple methods of data clustering. It works by dividing the dataset into a number of groups (clusters) based on the centroid value, which is the central point of each cluster. The user first determines the number of clusters (k), then the algorithm will place the data into the cluster that has the nearest centroid [9] .
K-Means belongs to the partitioning-based clustering method , where data is divided into exclusive groups, so that each data can only fit into one cluster. The basic principle is that each cluster has at least one data and each data must be on one specific cluster.
The quality of the clustering results is influenced by the similarity measure used and its ability to find hidden patterns in the data. A good cluster has a high internal similarity (intra-cluster similarity) and a large inter-cluster dissimilarity.
The following is a method of measuring distance on k-means using Euclidean Distance:

a. Euclidean Distance
Euclidean Distance is a distance measurement method that calculates the shortest (straight-line distance) between two points in a dimensional space. The formula is:

$$D_{(ij)} = \sqrt{\left(x_{1i} - x_{ij}\right)^2 + (x_{1i} - x_{2j})^2 + (x_{3i} - x_{3j})^2} \quad (1)$$

Information:
D_{ij} = Euclidean distance between the ith data and the jth centroid cluster
(x1i, x2i, x3i) = Variable value of the ith data
(x1j, x2j, x3j) = Centroid value on the variable for the jth cluster

This method is suitable for continuous data and takes into account the magnitude of the differences in each attribute. However, Euclidean Distance is quite sensitive to extreme differences in one variable, as the difference will be further magnified by quadratic operations.

3.5. Evaluation of Clustering Results

The evaluation of clustering results aims to assess the quality and consistency of the groups formed after the data grouping process. One of the methods used is variance calculation, which describes the rate of data spread within each cluster. Variance values can be used to measure how homogeneous the data is in a cluster and help compare the quality between clusters. The following is the formula for calculating the variance for each cluster:

Formula: $\text{Variance}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_i - c_k)$
Information:
nx is the number of data points in a cluster
xi is the i data point in the cluster
CK is the centroid of the cluster

1. Vminimum (Minumum Variance)
The minimum or minimum variance is the minimum variance of all clusters contained in a dataset. The process starts by calculating the variance for each cluster first, then taking the minimum variance value among all the clusters. Formulas: Vossium min(Variance, Variancez,..., Variancex) (2)

2. Vmaximum (Maximum Variance)
The maximum or maximum variance is the maximum variance of all the clusters contained in a data set. The process begins by calculating the variance for each cluster first, then taking the value of the myaxial variance among all the clusters.
Formula; Maximum V = max(Variance1, Variance2, Variance) (3)

3. Cluster Variance
Cluster variance is the average value of all cluster variance in a data set. This value provides a comprehensive overview of how widely distributed the data points are in the cluster [10] .
Rumus: Cluster Variance = $\frac{1}{n_k} \sum_{k=1}^{K} \text{Variance}_k$ (4)

4. Result and Discussion

The steps taken for the calculation of the grouping of extracurricular interest data of students of SMP Negeri 5 Binjai City use the clustering method, so that new information and knowledge can be produced for students and the agency can follow up on the problems that exist in

the inequality of student interest in extracurriculars, regarding how much extracurricular interest data of SMP Negeri 5 Binjai City students based on the type of extracurricular, activeness and performance.

4.1. Data Input

The input data in the system is in the form of data obtained from SMP Negeri 5 Binjai City. The input data will be transformed based on the transformation value of each variable used, the following is an explanation of the data that can be input and the variables and values of the transformation that are performed.

1. Input data
   Amount of data : 598 data
   Variable : X => Extracurricular Type, Y => Activity, Z => Performance
2. Cluster Grouping : 3 Cluster
3. Transformation value :

Table 1: Value of Data Transformation

Yes	Variable	Transformation	Value Transformation
1	Types of Extracurriculars	Roses	1
		Dance Arts	2
		Futsal	3
		Paskibra	4
		Scout	5
2	Activeness	Inactive	1
		Less Active	2
		Active	3
		Highly Active	4
3	Achievement	Not Participating in the Competition	1
		Participatory	2
		Winning School Level Competitions	3
		Winning City Level Competitions	4
		Winning the Provincial Level Competition	5

4.2. Grouping is carried out by processing data using 5 *clusters*.

1. Grouping of 3 clusters

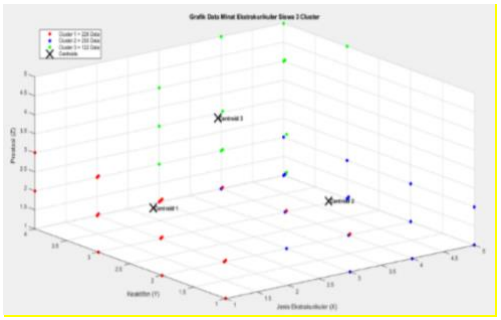

Fig. 1: Graph 3 Cluster

2. Grouping of 4 clusters

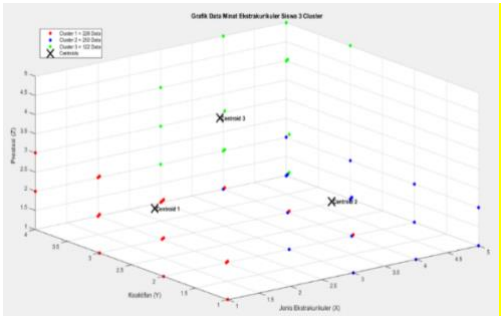

Fig. 2: Graph 4 Cluster

3. Grouping of 5 clusters

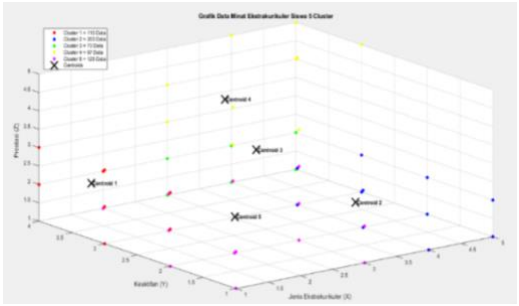

Fig. 3 : Graph 5 Clusters

Based on the image above, it can be seen that:

a. Cluster 1 includes 110 students who mostly participate in spiritual extracurricular activities, the dominant activeness in this type of extracurricular is active and the achievements achieved, namely participatory. This cluster reflects a group of students who are active in spiritual activities with a good level of participation in the activities they follow.

b. Cluster 2 includes 203 students who mostly participate in post-curricular extracurricular activities, the dominant activeness in this type of extracurricular is lack of activity and the achievements achieved, namely not participating in competitions. This cluster shows a group of students who participate in paskibra extracurriculars with a less active level of activity and are not involved in competitions.

c. Cluster 3 includes 73 students who mostly participate in post-curricular extracurricular activities, the dominant activity in this type of extracurricular is very active with the achievements achieved, namely participatory. This cluster represents students who are very active in paskibra with consistent involvement in activities despite not having achieved competitive achievements.

d. Cluster 4 includes 87 students who mostly participate in futsal extracurricular activities, the dominant activity in this type of extracurricular is very active with the achievements achieved, namely winning city level competitions. This cluster describes students who are very active in futsal extracurriculars with achievements at the city level.

e. Cluster 5 includes 125 students who mostly participate in dance extracurricular activities, the dominant activity in this type of extracurricular is less active with the achievements achieved, namely participatory. This cluster shows groups of students who participate in dance extracurriculars with a relatively low level of activity but still participate in activities.

The following is Table 2 which shows the centroid values of each variable (X, Y, Z) for the data grouping of 3, 4, and 5 clusters along with the amount of data in each cluster:

Table 2 : Grouping 3,4 and 5 Clusters

Cluster	*Centroid*	Variable X	Y	Z	Amount of Data
3	1	1,7876	2,9115	1,9381	226
	2	4,492	2,792	1,244	250
	3	3,5574	3,623	3,2459	122
4	1	3	2,9487	2,5726	117
	2	4,442	2,9058	1,3116	276
	3	4,3871	3,8387	4,7419	31
	4	1,4253	3,0575	1,9253	174
5	1	1,1455	3,3455	2,3818	110
	2	4,4631	2,5665	1,1724	203
	3	4,3836	4	1,7534	73
	4	3,4943	3,6092	3,6782	87
	5	2,44	2,424	1,576	125

4.3. Test Results Using Variance

The following are the test results of the 5 clusters:

1. Cluster 1 (1.1; 3.3; 2.6) with 110 data

Variance C1 = $\frac{1}{110}$ (0.2863 + 0.8317 + 1.3043 + ... + 0.5953) = 1.0410
Vmin = min (0.2863; 0.8317; 1.3043; ...; 0.59533) = 0.2863
Vmaks = max (0.2863; 0.8317; 1.3043; ...; 0.59533) = 2.2137

2. Cluster 2 (4.5; 2.5; 1.2) with 203 data

Variance C2 = $\frac{1}{203}$ (1.611 + 0.5059 + 0.4321 + ... + 0.5651) = 0.7650
Vmin = min (1.611; 0.5059; 0.4321; ...; 0.5651) = 0.4321
Vmaks = max (1.611; 0.5059; 0.4321; ...; 0.5651) = 3.4271

3. Cluster 3 (4.4; 4; 1.8) with 73 data

Variance C3 = $\frac{1}{73}$ (1.9752 + 0.4408 + 0.7148 + ... + 1.9752) = 0.5592
Vmin = min (1.9752; 0.4408; 0.7148; ...; 1.9752) = 0.2080
Vmaks = max (1.9752; 0.4408; 0.7148; ...; 1.9752) = 2.4820

4. Cluster 4 (3.5; 3.6; 3.7) with 87 data

Variance C3 = $\frac{1}{87}$ (0.7190 + 0.5006 + 1.7190 + ... + 0.8570) = 1.6488
Vmin = min(0.7190; 0.5006; 1.7190; ...; 0.8570) = 0.5006
Vmaks = max (0.7190; 0.5006; 1.7190; ...; 0.8570) = 4.3854

5. Cluster 3 (2.4; 2.4; 1.6) with 125 data

Variance C3 = $\frac{1}{125}$ (0.6732 + 4.4332 + 0.8252 + ... + 0.7052) = 1.1988
Vmin = min (0.6732; 4.4332; 0.8252; ...; 0.7052) = 0.6732
Vmaks = max (0.6732; 4.4332; 0.8252; ...; 0.7052) = 4.4332
Cluster variance = $\frac{1}{5}$ (1.0410 + 0.7650 + 0.5592 +1.6488 +1.1988) = 1.0426

The following are the test results of clusters 3,4 and 5 in Table 3:

Table 3 : Cluster Results Testing

Cluster	Centroid	Variance	Vmin	Vmaks	Cluster Variance
3	1,7; 2,9; 1,9	1.8282	0.0568	5.1542	1.6147
	4,4; 2,7; 1,2	1.1031	0.3449	5.4969	
	3,5; 3.6; 3,2	1.9129	0.5133	5.5461	
4	3; 2,9; 2,5	0.9429	0.1853	4.1253	1.4049
	4,4; 2,9; 1,3	2.1810	0.3013	5.8085	
	4,3; 3,8; 4,7	0.8866	0.2425	2.5005	
	1,4; 3; 1,9	1.6091	0.1898	5.2704	
5	1,1; 3,3; 2,3	1.0410	0.2863	2.2137	1.0426
	4,4; 2,5; 1,1	0.7650	0.4321	3.4271	
	4,3; 4; 1,7	0.5592	0.2080	2.4820	
	3,4; 3,6; 3,6	1.6488	0.5006	4.3854	
	2,4; 2,4; 1,5	1.1988	0.6732	4.4332	

From the table above, it can be deduced:

1. In cluster 5
    a. Grouping with 5 clusters yields the lowest cluster variance of 1.0426, which indicates that the data distribution between clusters is getting more compact. The first cluster has a variance of 1.0410, with a Vmin of 0.2863 and a Vmax of 2.2137, indicating a relatively small spread.
    b. The second cluster recorded a variance of 0.7650, with a Vmin of 0.4321 and a Vmax of 3.4271, indicating that most of the data is close enough to the center of the cluster.
    c. The third cluster shows a variance of 0.5592, with a Vmin of 0.2080 and a Vmax of 2.4820, reflecting a good degree of compactness.
    d. The fourth cluster has a variance of 1.6488, with a Vmin of 0.5006 which is quite high, indicating data proximity, but Vmax 4.3854 still indicates the existence of distant data.
    e. The fifth cluster shows a variance of 1.1988 with a high Vmin of 0.6732 which shows the data very close to the center of the cluster, although the Vmax of 4.4332 is still worth watching.

    Overall, the results of clustering 5 groups showed the best results because the value of variance between clusters was lowest compared to 3 and 4 clusters, indicating that the data was increasingly well grouped.

5. Conclusion

Based on the research that has been conducted to group the extracurricular interests of SMP Negeri 5 Binjai City students using the k-means algorithm, several conclusions can be drawn as follows:

1. This study succeeded in grouping the interests of students of SMP Negeri 5 Binjai City using the k-means algorithm which was successfully applied through the Matlab R2014b programming application for the process of grouping extracurricular interest data based on the variables of extracurricular type, activeness and achievement achieved by students with three scenarios of the number of clusters, namely 3, 4 and 5 clusters.
2. The results of the analysis of student interest patterns showed that three clusters distinguished students with high, medium, and low activeness. The four clusters show a more specific pattern with a separation between active students who have not achieved and those who have achieved achievements at the school or city level. The five clusters resulted in the most detailed segmentation with small groups of high-achieving students at the city or provincial level that were separate from low-achieving active students. The more clusters, the more detailed the pattern is but the uniformity within each cluster decreases.
3. Based on the evaluation, clustering can be used to support the management and development of extracurricular activities at SMP Negeri 5 Binjai City by adjusting the number of clusters to the purpose of the analysis. Three clusters are sufficient for general strategy, while four or five clusters provide more detailed information for specific coaching. These results help schools set strategies, allocate tutors and facilities, and design motivational programs for low-activity students to increase participation.

References

[1] T. Alivia, U. Sultan Aji Muhammad Idris Samarinda, U. Sultan Aji Muhammad Idris Samarinda Jl HAM Rifatddin, K. Loajanan illir, K. samarinda, and K. Timur, "Character Education Management Through Extracurricular Activities," 2023.

[2] I. Gede, I. Sudipa, and M. Darmawiguna, "DATA MINING TEXTBOOK," 2024. [Online]. Available: https://www.researchgate.net/publication/377415198

[3] B. Laksono, Y. Syahidin, and Y. Yunengsih, "Implementation of Data Mining for Inpatient Data Clustering with K-Means Clustering Algorithm," *Journal of Information Systems and Application Technology*, vol. 7, no. 2, pp. 621–627, Apr. 2024, doi: 10.32493/jtsi.v7i2.39354.

[4] M. Putri Chandra and M. Alridho Lubis, "Efforts to Increase Students' Learning Interest by Implementing Information Services Using Post-Covid Audio-Visual Media," 2023.

[5] D. P. Br. Marpaung, Nurroyian, Hasbih Sholeh Suryadi, Lucky Tirta Ardiansyah, and Muhammad Iqbal, "The Role of Extracurricular Activities in the Development of Students' Social Skills," *Indo-MathEdu Intellectuals Journal*, vol. 5, no. 3, pp. 3408–3416, Jul. 2024, doi: 10.54373/imeij.v5i3.1365.

[6] A. Tjolleng, "Introduction to MATLAB programming: A practical guide to learning MATLAB," 2017. [Online]. Available: https://www.researchgate.net/publication/334945947

[7] D. Papakyriakou and I. S. Barbounakis, "Data Mining Methods: A Review," *Int J Comput Appl*, vol. 183, no. 48, pp. 5–19, Jan. 2022, doi: 10.5120/ijca2022921884.

[8] P. T. I. U. P. Y. Nurirwan Saputra, "Introduction to Data Mining," 2023.

[9] Khairul Solekhan Arif, "CLUSTERING DATA OF PUHPELEM PUSKESMAS PATIENTS USING K-MEANS CLUSTERING," 2023.

[10] Dwi Astuti, Relita Buaton, and Magdalena Simanjuntak, "Clustering of Biological Food Poisoning Case Data Based on Causal Factors Using the Clustering Method," *Bridge: Journal of Information and Telecommunication Systems Publications*, vol. 2, no. 4, pp. 19–31, Aug. 2024, doi: 10.62951/bridge.v2i4.199.