Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Improving the Regional Grouping Model for Students of SMK Muthia Harapan Using K-Means Clustering Algorithm

Salma Nur Fikriani[1]*, Ade Irma Purnamasari[2], Agus Bahtiar[3], Edi Wahyudin[4]

[1,2,3,4] *STMIK IKMI Cirebon*

nf.salmaama@gmail.com[1]*, irma2974@yahoo.com[2], agusbahtiar038@gmail.com[3], ediwahyudin@gmail.com[4]

Abstract

Education is an important aspect in human life to improve and develop self-potential. The rapid development of technology has increased the need for fast, accurate, and efficients information, including in the world of education. One of the challenges faced by SMK Muthia Harapan Cicalengka is the accumulation of student data every year. This makes it difficult to identify student data based on region of origin. This research aims to apply data mining using the K-Means Clustering method to group student data with similar characteristics. The method used in this research is Knowledge Discovery in Database (KDD) which includes the stages of data cleaning, data transformation, data mining, and evaluation. The implementation og K-Means Clustering is done using RapidMiner with attributes such as Name, Village, Department, and school of origin. The purpose of this research is to provide a targeted and strategic overview of areas that can have a significant impact on the supply of students each year. The result show that student data can be grouped into two clusters. Cluster 0 consists of 254 items and cluster 1 consists of 254 items, with a Davies-Bouldin Index (DBI) value of 0.549.

Keywords: *Data Mining, K-Means, Knowledge Discovery in Database, Student Home Regions, SMK Muthia Harapan Cicalengka*

1. Introduction

The progress of science is currently very developed and can be seen in all fields such as industry, business, and education. Education is a conscious and planned effort to create a conducive learning atmosphere and student learning interactions effectively foster their ability to have religious spiritual strength, calmness, character, knowledge, and abilities needed for themselves, society, nation, and state. Given the purpose of this education, the quality and implementation of the teaching and learning process in schools or in educational institutions must be improved. education occupies a strategic position in human resource development, especially at the level of vocational secondary education designed to prepare a skilled workforce. [1]

SMK Muthia Harapan Cicalengka as one of the vocational education institutions experiences challenges in attracting students who match the characteristics and development needs of the institution. One of the significant challenges faced by the school is in determining the area of origin of students who have the potential to be targeted for promotion more effectively. data-based promotion strategies are becoming increasingly relevant along with the rapid development of information technology. The use of K-Means Clustering algorithm as a tool in data clustering has been proven to provide accurate results in various fields, including education. The K-Means Clustering algorithm is an alternative in helping educational institutions segment areas based on student potential that can be obtained from the student's area of origin, which can help schools determine more targeted and efficient marketing strategies.

The main objective of this research is to develop a model of mapping the area of origin of students of SMK Muthia Harapan Cicalengka using the K-Means Clustering Algorithm. With this modeling, this research seeks to produce accurate and relevant information about the distribution patterns of students' areas of origin, which can be the basis for designing more effective and targeted promotional strategies. This research also aims to fill the knowledge gap in the application of data mining in the field of education, especially in the context of geographic data mapping of students at the Vocational High School level. The expected contribution of this research lies not only in increasing understanding in cluster analysis, but also in providing practical insights for educational institutions related to optimizing marketing strategies. In addition, this research can be an important reference in the application of K-Means for educational data management.

2. Literature

The results of the literature review that has been carried out in research journals related to the topic “K-Means Clustering Algorithm to improve the student origin region clustering model” can be described as follows:

The paper [2] discusses the K-Means Clustering algorithm to process student data and extract useful information for strategic decision making. The goal is to use the K-Means Clustering algorithm to analyze student data and provide a targeted and strategic overview of provinces that have a significant impact on student supply each year. The results show that the largest supply of students comes from central Java, while the smallest comes from bangka belitung, west sulawesi, banten, east kalimantan, west papua, and bengkulu.

Paper [3] discusses the problems experienced by the study center course in the city of kisaran which has difficulty in determining the location of the right and fast promotion of new student admissions. This is because the promotions that have been carried out sometimes encounter obstacles. In addition, the process of determining a good promotional location requires research and consideration of many aspects, so it takes a lot of time. The researcher proposes the application of data mining using the K-Means Clustering method to group prospective student data based on similar characteristics. The goal is to obtain information about areas with high potential to bring in new students so that decision making in determining promotional locations can be done effectively and efficiently.

Paper [4] discusses the application of the K-Means Clustering algorithm to analyze and group new prospective student data at STMIK Primakarta, with the CRISP-DM (Cross-Industry Standard Process for Data Mining) method, giving the result that 3 clusters of prospective students are formed with their respective characteristics. Cluster 1 consists of 906 students, mostly from the East Denpasar and Gianyar regions. Cluster 2 consists of 28 students from distant areas, and cluster 3 consists of 77 students also from distant areas. Based on these results, a more effective promotion strategy can be formulated, such as conducting direct promotion, increasing promotion to schools, marketing to schools in the same major, and providing additional discounts in the first wave. For distant areas, promotion can be done through advertising.

Paper [5] discusses the application of the K-Means Clustering method to determine promotional strategies at the Assholeh Pemalang College of Economics. The method used is clustering with the K-Means algorithm, the data analysis process uses Knowledge Discovery in Database (KDD) techniques and the implementation uses WEKA software version 3.8.5. The implementation process resulted in 3 clusters, namely cluster 1 (42% of data), cluster 2 (41% of data), and cluster 3 (17% of data). So that the suggested promotion strategy is to send the marketing team to the dominating areas, conduct socialization, distribute brochures, and align with the promotion strategy [6], [7], [8].

3. Research Methods

This research uses a quantitative descriptive approach. In its understanding, descriptive research is carried out by seeking information related to existing symptoms, planning how to approach it, clearly explaining the objectives to be achieved and collecting various kinds of data as material for making reports.

3.1. Research Stages

As for analyzing data in the application of data mining, it uses the Knowledge Discovery in Database (KDD) stage process which consists of data, data cleaning, data transformation, data mining, and pattern evolution.

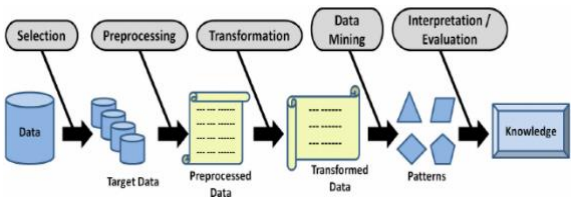

Figure 1: Research Method

Below is a discussion of the knowledge discovery in database (KDD) stage:
1. Data selection; The data selected is student data for SMK Muthia Harapan for the 2021-2024 school year.
2. Data preprocessing; Perform data cleaning by cleaning data from missing or inconsistent attributes to ensure high data quality.
3. Data transformation; Converting the selected data to make it suitable for the data mining process.
4. Data mining; Perform data clustering using the K-Means algorithm and RapidMiner application to assist analysis.
5. Evaluation; Evaluate the clustering results to assess the relevance and accuracy of the clustering results.

3.2. Data Source

In this study using primary and secondary data obtained directly from the first source or not through intermediaries. The primary data generated is a dataset of biodata of 10th to 12th grade students at SMK Muthia Harapan Cicalengka. The dataset consists of 508 student data with 4 attributes, namely Name, Village, Department, and School Origin. While the secondary data used is information obtained from literature such as journals, containing explanations and groupings of the K-Means algorithm.

4. Result and Discussion

4.1. Data Selection

The data to be processed in this research refers to the dataset of student personal data of SMK Muthia Harapan Cicalengka. This dataset has 508 items with 4 attributes including Name, Village, Department and School of Origin.

Table 1: data to be processed into RapidMiner

NO	NAME	NEIGHBORHOOD	MAJOR	SCHOOL ORIGIN
1	AA SUTISNA	MANDALAWANGI	XII TBSM 2	SMPN 2 CICALENGKA
2	ABDI FARIS SANJAYA	NAGROG	XII RPL 2	SMPN 2 CICALENGKA
3	ABDUL MARIX MAULANA	BABAKAN PETEUY	X RPL 2	SMPN 1 CICALENGKA
4	BROR RIZKY	CIKUYA	XII TBSM 1	SMPN 1 CICALENGKA
5	ADE YUSRIL SAPUTRA	MARGAASIH	XII RPL 2	SMPN 2 CICALENGKA
6	ADELIA PUTRI	BABAKAN PETEUY	X RPL 3	SMPN 2 CICALENGKA
7	ADI AHMAD ABDUL ROUF	\HEGARMANAG	XI TBSM 1	MTS AL-HIDAYAH PANGUYUBAN
8	ADI PERMANA	MARGAASIH	X TBSM 2	SMP YADIKA 1 CICALENGKA
9	ADITYA JUNIAWAN SIDIK	SINDANGPAKUON	X RPL 4	SMP YADIKA 1 CICALENGKA
…..	…..	…..	…..	…..
505	YUSLI	GANJAR SABAR	X11 TBSM 3	SMPN 1 NAGREG
506	ZAENAL FADLAN	TENJOLAYA	XI TBSM 3	SMP PLUS GANESHA
507	ZAHRA NUR ARIFA	BABAKAN PETEUY	X RPL 2	SMP PASUNDAN 12
508	ZIDAN ZULKAHFI	BABAKAN PETEUY	X TBSM 1	SMP PASUNDAN 12

The first step in data selection involves the use of the read excel operator. The function of this operator is to read the information related to the grouping of student data by region of origin stored in the MS-Excel file.

Figure 2: Excel Read Operator

After using the read excel operator, then use the set role operator to change the role of the attribute, and it will be included in the Id category.

Figure 3: Operator Set Role

4.2. Preprocessing Data

Data preprocessing is done by checking empty data, incomplete and missing data will be deleted, because the presence of irrelevant data can reduce the level of accuracy in the data mining process. The operator used in this stage is the Missing Values Operator.

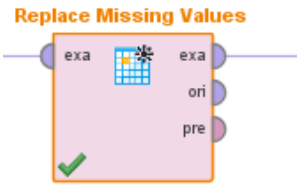

Figure 4: Operator Replace Missing Value

4.3. Transformation Data

Data transformation is the process of changing or modifying data to make it more suitable and effective for analysis. The goal is to make the data easier to analyze and fit the model used.

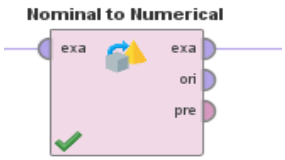

Figure 5: Operator Nominal to Numerical

4.4. Data Mining

In the data mining process, the method used in clustering students' home regions uses K-Means Clustering while the evaluation process for performance testing uses the cluster distance performance operator.

Figure 6: Operator K-Means Clustering

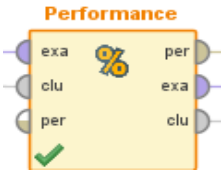

Figure 7: Operator Cluster Distance Performance

After various stages of careful performance testing, including data analysis, effectiveness measurement, and in-depth evaluation of key parameters, an average result (avg) was obtained that reflects the overall efficiency and effectiveness of the process.

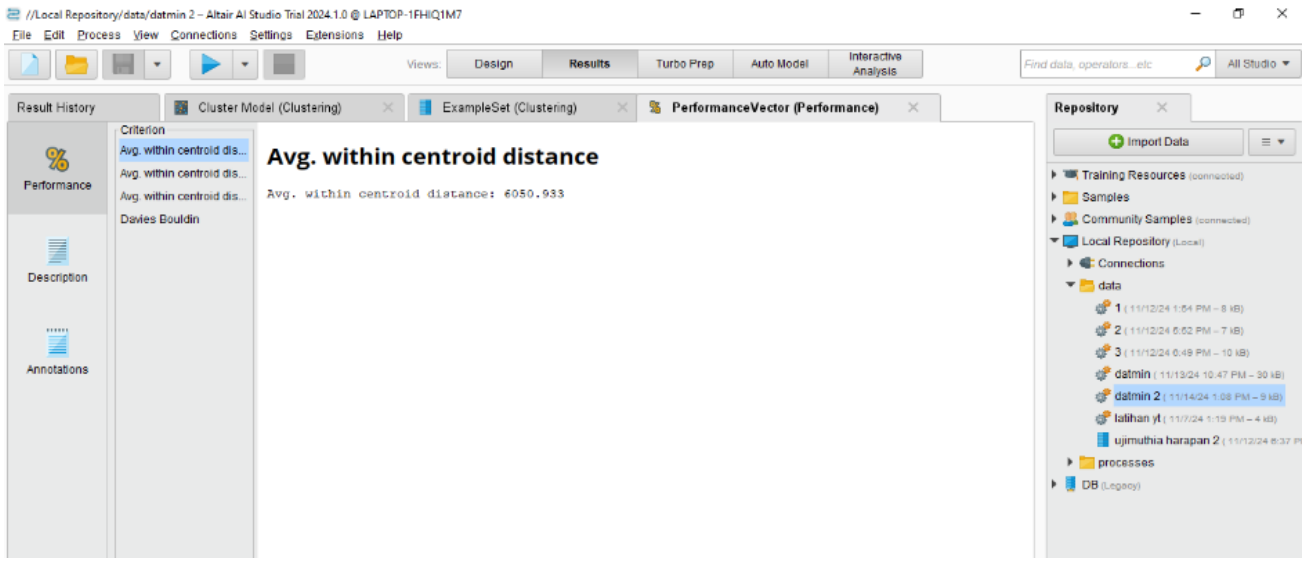

Figure 8: Avg Search Result

4.5. Evaluation

In model evaluation, the information is converted into a simpler format for easier understanding. This process is done to get the best clusters by looking at the results of the Davies Bouldin Index (DBI) evaluation. DBI is a metric that measures how well clusters are formed, with lower values indicating better clustering.

This study tested various K values ranging from 2 to 10 clusters, to evaluate the clustering performance based on the evaluation results of the lowest DBI value at K = 2 with a value of 0.549.

Table 2: K-Value Trial Results

Kluster	DBI
2	0.549
3	0.608
4	0.662
5	0.731
6	0.782
7	0.769
8	0.730
9	0.738
10	0.691

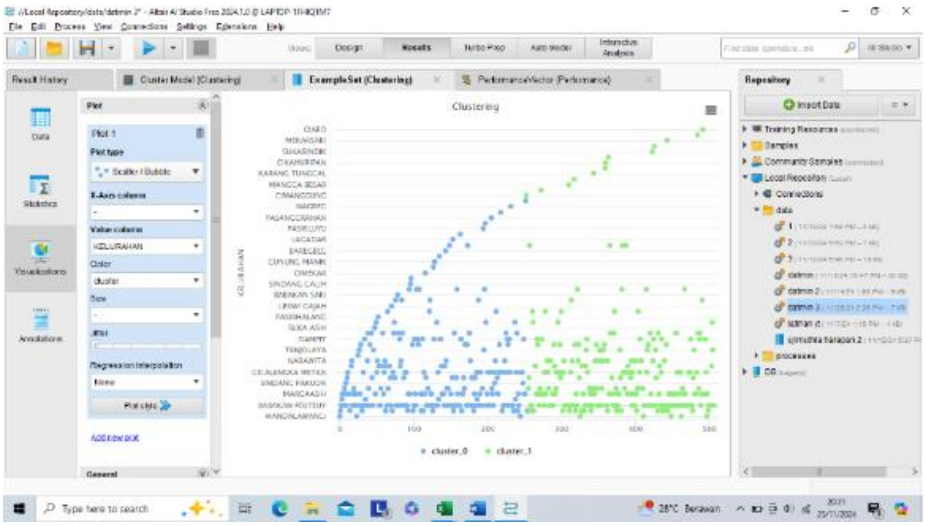

Figure 9: Graph Plot

Based on the results of the graph plot above, it can be seen that the majority of SMK Muthia Harapan Cicalengka students come from Babakan Peuteuy, Margaasih, and Cicalengka Wetan villages. While the minority of students come from the Ciaro, Nagreg, and Lagadar villages.

5. Conclusion

Based on the discussion that has been described, the researcher gets a conclusion that the K-Means Clustering Algorithm method can be applied in grouping the regions of origin of students of SMK Muthia Harapan Cicalengka. The analyzed data resulted in two clusters with a balanced number, each cluster numbering 254 items and getting a DBI value of 0.549. Babakan peteuy, Margaasih and Cicalengka Wetan are the areas with the highest number of students in both clusters, which reflects the high concentration of students from these areas.

Reference

[1] A. Bellanov and L. Nurhayati, "K-Means Clustering Analysis Untuk Menentukan Strategi Promosi Kampus," *J. Tek. Ind. J. Has. Penelit. dan Karya Ilm. dalam Bid. Tek. Ind.*, vol. 9, no. 1, p. 259, 2023, doi: 10.24014/jti.v9i1.22492.

[2] R. L. Pattipeilohy and M. A. I. Pakereng, "Penerapan K-Means Clustering Pada Data Mahasiswa Fakultas Interdisiplin Program Studi D4 Destinasi Pariwisata Untuk Menentukan Strategi Promosi," *J. Sains Komput. Inform. (J-SAKTI*, vol. 7, no. 1, pp. 320–331, 2023.

[3] N. Azmi, F. Helmiah, and S. Sudarmin, "Implementasi Metode K-Means Sebagai Upaya Penentuan Lokasi Promosi Penerimaan Siswa Baru," *Build. Informatics, Technol. Sci.*, vol. 3, no. 4, pp. 649–660, 2022, doi: 10.47065/bits.v3i4.1456.

[4] Oki Oktaviarna Tensao, I Nyoman Yudi Anggara Wijaya, and Ketut Queena Fredlina, "Analisa Data Mining dengan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi Mahasiswa Baru Pada STMIK Primakara," *Inf. (Jurnal Inform. dan Sist. Informasi)*, vol. 14, no. 1, pp. 1–17, 2022, doi: 10.37424/informasi.v14i1.135.

[5] N. A. Rahmalinda and A. Jananto, "Penerapan Metode K-Means Clustering Dalam Menentukan Strategi Promosi Berdasarkan Data Penerimaan Mahasiswa Baru," *J. Tekno Kompak*, vol. 16, no. 2, p. 163, 2022, doi: 10.33365/jtk.v16i2.1971.

[6] Pardede, A. M. H. (2019). Metode K-Means untuk pengelompokan masyarakat miskin dengan menggunakan jarak kedekatan Manhattan City Dan Euclidean (Studi kasus kota binjai). *Journal Information System Development (ISD)*, *4*(2).

[7] Syahputra, S., Ramadani, S., & Pardede, A. M. H. (2020). Menentukan Strategi Promosi Menggunakan Algoritma Clustering K-Means. *JOISIE (Journal Of Information Systems And Informatics Engineering)*, *4*(1), 7-14.

[8] Arbaeti, E. E., Pardede, A. M. H., & Kadim, L. A. N. (2023). Application of K-Means Clustering Algorithm to Analyze Insurance Company Business (Case Study: Pt. Jasindo Insurance). *Journal of Mathematics and Technology (MATECH)*, *2*(2), 173-192.