Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

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

K-Means Clustering to Improve Interest Grouping Model For High School Students

Dewi Rengganis[1]*, Ade Irma Purnamasari[2], Agus Bahtiar[3], Edi Tohidi[4]

[1,2]*Informatics Engineering, STMIK IKMI Cirebon*
[3]*Information Systems, STMIK IKMI Cirebon*
[4]*Computerized accounting, STMIK IKMI Cirebon*
dewirengganis218@gmail.com[1*], irma2974@yahoo.com[2], agusbahtiar038@gmail.com[3], editohidi00@gmail.com[4]

Abstract

Informatics Engineering has a significant appeal to high school students in the digital era. However, differences in students' understanding of career prospects in this field affect their level of interest. This study aims to identify students' interest patterns using the K-Means Clustering algorithm as a basis for developing data-based strategies to increase the attractiveness of the major. This study used quantitative methods with primary data collected through questionnaires from 202 high school students. The variables analyzed include students' understanding of Informatics Engineering, interest in technology subjects, and aspirations to continue their studies in the field. The data was processed using RapidMiner software, through the stages of pre-processing, data transformation, and model evaluation. Davies-Bouldin Index (DBI) was used to determine the best number of clusters, with cluster trials (k) from 2 to 10. The results showed the best DBI value at k=2 with a score of 0.527. Two clusters were formed: Cluster 0 (uninterested students) with 96 students and Cluster 1 (interested students) with 106 students. Interested students generally have a better understanding of career prospects in technology, while less interested students need additional education to increase their interest. This research shows the importance of a data-driven approach in understanding student needs. For students with low interest, an upstream program is needed.

Keywords: *Student interest, K-Means clustering, Educations, Informatics, Engineer, Davies Bouldin Index*

1. Introduction

Education is one aspect of national development that is very important to realize the development of human resources and national character. Therefore, children's education from an early age is very important to increase children's intelligence which is the basis for further self-development to get the best value in the teaching and learning process[6]. The choice of major is a significant decision for students because it can determine the direction of their education and career in the future. One of the majors that are in great demand by high school students is Informatics Engineering, which offers promising career prospects in today's digital era. So it is not surprising that we need to analyze their interests in this case for a more planned future [7]. The majority of students who are confused and have difficulty determining their interests end up falling for trends or promotions with the lure of easy majors that pass by on social media without considering the consequences or dependents that will become their burden later. This becomes one of the problem factors that later have an impact on their inconsistency in what they choose, then end up being burdened because they take a major that they are not really interested in and enjoy in the process[2].

Grouping students' interest in learning mathematics using the K-Means Clustering Algorithm is easier to partition data into groups, so that the same data can be grouped in the same cluster. then there is quite related research, namely[1], [5] using the same method successfully reported the mapping of high, medium to lowest student achievement[3]. Reporting the results of how the K-Means Clustering Algorithm method can prove the results of the study quite accurately and of course can be used as a reference for decision making to choose majors that are in demand and not in demand[4].

Along with the development of technology, the use of algorithms in analyzing data has become very relevant. One algorithm that is widely used in data analysis is the K-Means Clustering algorithm, which aims to group data into several groups based

on the similarity of their characteristics. In this context, the K-Means algorithm can be used to analyze and identify patterns of student interest in Informatics Engineering majors, based on a number of relevant variables. this research was conducted to at least know and analyze potential successors in the field of technology earlier.[8]
This research was conducted to categorize the interest of final high school students towards Informatics Engineering majors based on several general characteristics.

2. Research Methodology

This research uses the K-Means Clustering Algorithm method, with a quantitative approach. The quantitative approach was chosen because it allows researchers to measure variables objectively and produce data that can be analyzed statistically. This research uses a primary approach and collection method, namely by filling out questionnaires for around 202 12th grade students of PGRI Cicalengka High School.
The stages of the Knowledge Discovery in Database (KDD) process used in data analysis in this data mining include data selection, data preprocessing, data transformation, data mining process, and evaluation. Samples were taken using cluster sampling technique. This technique was chosen to ensure that the sample represents the various backgrounds that exist in the population, such as classes and majors taken. The questionnaire consisted of 4 question items covering personal data, general questions, interests and their knowledge of the Informatics Engineering major. Data was collected through the distribution of questionnaires given to students directly.
The data obtained will be analyzed using Rapidminer software, which allows efficient data processing. The results of the analysis will be presented in the form of tables and graphs for easy interpretation and visualization of the data.
Planning and data collection will be carried out simultaneously around the time period of november 2024.

3. Result and Discussion

The stages of the Knowledge Discovery in Database (KDD) process used in data analysis in this data mining include data selection, data preprocessing, data transformation, data mining process, and evaluation.

3.1. Research Result

3.1.1. Data Selection

The data used in this study are questionnaire data on the interests of PGRI Cicalengka High School students from November 13, 2024 to November 29, 2024 with a total dataset of 202, and several attributes including name, class, gender, knowing about informatics engineering majors, liking subjects related to computers and technology, and interest in continuing education in informatics majors. By converting some categorical variables into numeric.

Such as converting the categories of not interested, interested and very interested into numbers. For example, not interested=0 and interested=1.

Table 1: Data Selection

NO	NAMA	KELAS	JENIS KELAMIN	Mengetahui tentang Jurusan Teknik Informatika	Menyukai mata pelajaran yang berkaitan dengan komputer dan Teknologi	Minat untuk melanjutkan pendidikan di jurusan Teknik Informatika
1	nur'aina dzulqa	12 IPA 1	PEREMPUAN	tahu	Suka	Minat
2	Nur Mardiah	12 IPA 1	PEREMPUAN	tahu	Tidak suka	Minat
3	nurul halimah agustina	12 IPA 1	PEREMPUAN	Tahu	Suka	minat
4	Risya nurfitri	12 IPA 1	PEREMPUAN	Tahu	Tidak suka	Tidak minat
5	Nawaal Nabiilah	12 IPA 1	PEREMPUAN	Tahu	Suka	Minat
6	M.shavier.janeti	12 IPA 1	LAKI-LAKI	Tahu	Suka	Minat
7	Gina nuraeni	12 IPA 1	PEREMPUAN	tahu	Suka	Minat
8	KANIA RAMADHANI	12 IPS 1	PEREMPUAN	Tahu	Suka	Minat
	----	----	---	---	---	---
197	R. Luis Saefullah	12 IPS 4	LAKI-LAKI	Tahu	Suka	Minat
198	Jessica Pangestu	12 IPS 2	PEREMPUAN	Tahu	Suka	Tidak minat
199	Ridwan Hasbullah	12 IPS 2	LAKI-LAKI	Tahu	Tidak suka	Tidak minat
200	Sabri Najmudin	12 IPS 2	LAKI-LAKI	Tahu	Suka	Minat
201	Digdaya Adriansyah	12 IPS 2	LAKI-LAKI	Tahu	Tidak suka	Minat
202	Ophelia Saragih	12 IPS 2	LAKI-LAKI	Tidak tahu	Tidak suka	Tidak minat
203	Kamila Rajata	12 IPA 4	PEREMPUAN	Tidak tahu	Tidak suka	Tidak minat

3.1.2. Pre-Processing

After the data selection stage, the next stage is preprocessing. The purpose of pre-processing is to remove unnecessary data attributes, null or missing data, eliminate inconsistent data, and add id to the dataset to be used.

Row No.	NAMA	KELAS	JENIS KELA...	mengetahui ...	menyukai m...	minat untuk ...
1	nur'aina dzulqa	12 IPA 1	PEREMPUAN	1	1	1
2	Nur Mardiah	12 IPA 1	PEREMPUAN	1	0	1
3	nurul halima...	12 IPA 1	PEREMPUAN	1	1	1
4	Risya nurfitri	12 IPA 1	PEREMPUAN	1	0	0
5	Nawaal Nabil...	12 IPA 1	PEREMPUAN	1	1	1
6	M.shavier.jan...	12 IPA 1	LAKI-LAKI	1	1	1
7	Gina nuraeni	12 IPA 1	PEREMPUAN	1	1	1
8	ZHAFARINA F...	12 IPA 1	PEREMPUAN	1	0	0
9	Anggun Nur F...	12 IPA 1	PEREMPUAN	1	1	1
10	Cita Rosita	12 IPA 1	PEREMPUAN	1	1	0
11	Aliya Putri Ma...	12 IPA 1	PEREMPUAN	1	1	0
12	Siti Nurhasan...	12 IPA 1	PEREMPUAN	1	1	0
13	Anisa Ramad...	12 IPA 1	PEREMPUAN	1	1	0

ExampleSet (202 examples, 0 special attributes, 6 regular attributes)

Fig. 1: Data processing result

3.1.3. Transformation

At the transformation stage, the author changes the type of data attributes that were originally non-numeric into numeric data types. The data attributes that have been changed are class attributes, gender, knowing about informatics engineering majors, liking lessons related to computers and technology, and interest in continuing education in informatics engineering majors, into numbers using Nominal to Numerical to match the type of data needed by the K-Means algorithm in the clustering process.

Row No.	NAMA	KELAS	JENIS KELA...	mengetahui ...	menyukai m...	minat untuk ...
1	nur'aina dzulqa	0	0	1	1	1
2	Nur Mardiah	0	0	1	0	1
3	nurul halima...	0	0	1	1	1
4	Risya nurfitri	0	0	1	0	0
5	Nawaal Nabil...	1	0	1	1	1
6	M.shavier.jan...	0	1	1	1	1
7	Gina nuraeni	0	0	1	1	1
8	ZHAFARINA F...	1	0	1	0	0
9	Anggun Nur F...	0	0	1	1	1
10	Cita Rosita	0	0	1	1	0
11	Aliya Putri Ma...	0	0	1	1	0
12	Siti Nurhasan...	0	0	1	1	0

ExampleSet (202 examples, 1 special attribute, 5 regular attributes)

Fig. 2: Transformation results

3.1.4. Data Mining

After carrying out the transformation stage, the next stage is to apply to data mining. At the stage of applying data mining, the author uses rapidminer tools version 9.1.0 with the K-Means algorithm method. At the data mining stage, the author divides it into two stages, namely the first stage the author performs the clustering process using K-Means with the operator used, namely the Clustering operator (K-Means) and the second stage the author tests and evaluates the results of clustering using the Cluster Distance Performance operator with the David Bouldin Index (DBI) evaluation method.
K-Means Clustering Stages.
The first stage is the author performs the K-Means clustering process, experimenting from k
= 2 to k = 10. Data mining modeling on student specialization data clustering using the K-Means algorithm can be seen in Figure 3 below.

Fig. 3: K-Means Clustering

3.1.5. Evaluation

The next stage is testing the clustering results by applying the DBI value to the Cluster Distance Performance parameter. In DBI testing, the cluster that has the smallest DBI value or close to 0 is the best cluster. To find the best cluster, the author experiments from cluster 2 to 10 recapitulation. The number of clusters generated from the Davied Bouldin Index results can be seen from table 2 below.

Table 2: Evaluation DBI Results

Clustering	Number of Cluster Member	Hasil Nilai DBI
2	Cluster 0 : 96 items Cluster 1 : 106 items	0.527
3	Cluster 0 : 46 items Cluster 1 : 84 items Cluster 2 : 72 items	0.617
4	Cluster 0 : 63 items Cluster 1 : 61 items Cluster 2 : 44 items Cluster 3 : 34 items	
5	Cluster 0 : 34 items Cluster 1 : 34 items Cluster 2 : 60 items Cluster 3 : 24 items Cluster 4 : 50 items	0.747
6	Cluster 0 : 33 items Cluster 1 : 24 items Cluster 2 : 28 items Cluster 3 : 60 items Cluster 4 : 21 items Cluster 5 : 36 items	0.858
7	Cluster 0 : 36 items Cluster 1 : 38 items Cluster 2 : 28 items Cluster 3 : 23 items Cluster 4 : 33 Items Cluster 5 : 21 items Cluster 6 : 23 items	1.031
8	Cluster 0 : 24 items Cluster 1 : 25 items Cluster 2 : 30 items Cluster 3 : 34 items Cluster 4 : 21 items Cluster 5 : 9 items Cluster 6 : 26 items Cluster 7 : 33 items	1.139
9	Cluster 0 : 24 items Cluster 1 : 18 items Cluster 2 : 19 items Cluster 3 : 28 items Cluster 4 : 23 items Cluster 5 : 21 items Cluster 6 : 37 items Cluster 7 : 16 items Cluster 8 : 16 items	
10	Cluster 0 : 30 items Cluster 1 : 34 items Cluster 2 : 28 items Cluster 3 : 13 items Cluster 4 : 24 items Cluster 5 : 13 items Cluster 6 : 13 items Cluster 7 : 12 items Cluster 8 : 14 items Cluster 9 : 21 items	1.139

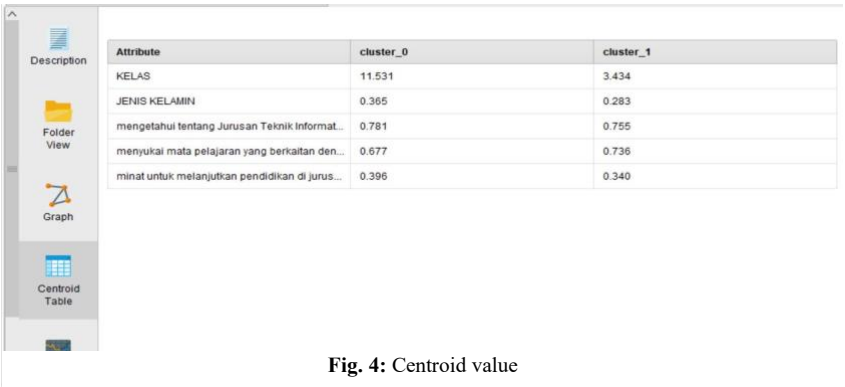

Fig. 4: Centroid value

At the evaluation stage is the stage in producing patterns to estimate the expected goals. Based on table 2 shows that the cluster data for the specialization of PGRI Cicalengka High School students using the Davies Bouldin Index calculation is the closest value to 0 with cluster 2 to cluster 10 experiments resulting in the best k value in cluster 2 which is 0.527 with the number of Cluster 0 members: 96 items, Cluster 1: 106 items.

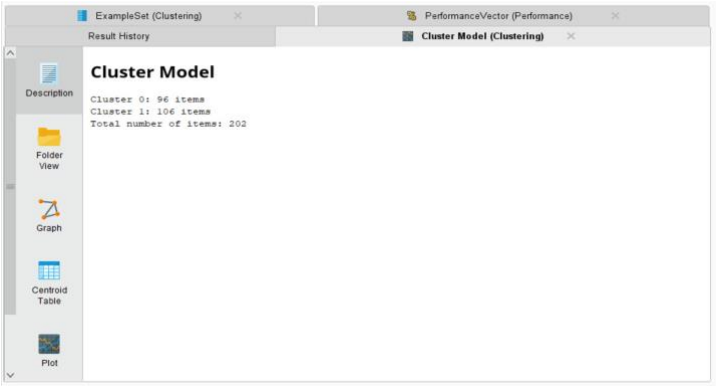

Fig. 5: Cluster results

The following is the distance between cluster centroids from the specialization data of PGRI Cicalengka High School students using the K-Means algorithm can be seen in Figure 5 and the results of the Performance Vector description can be seen in Figure 6.

Fig 6: Avg.within Performance

3.2. Discussion

The first stage of data mining is the author doing the K-Means clustering process, experimenting from k = 2 to k = 10.

Table 3: DBI result

Cluster	Nilai
2	0.527
3	0.617
4	0.645
5	0.747.
6	0.858
7	1.031
8	1.139
9	1.217
10	1.139

The results of the DBI value above can be concluded that the k value that is close to the optimum is k = 2 with a DBI value of 0.527. This shows that the most ideal data grouping is with two clusters, which allows the K-Means algorithm to produce a clearer pattern related to the level of student interest in the Informatics Engineering major.
After determining k=2 as the best number of clusters, an interpretation of the clusters formed was carried out. The first cluster (Cluster 0) includes students who are not interested in continuing their education in Informatics Engineering, while the second cluster (Cluster 1) includes students who have an interest in continuing their education in that field.
The K-Means Clustering algorithm utilizes RapidMiner tools to produce a table view of the centroid distance of each cluster as follows.

The following are the results of data processing using the K-Means method in the form of bar charts.

a. Grouping based on their interest in Informatics Engineering

Of the total student population obtained, 202 people. Divided into two clusters, Cluster 0 and Cluster 1. Cluster 0 is students who are not interested in majoring in Informatics Engineering and Cluster 1 who are interested in the major. In the graph, Cluster 0 has 96 items, and Cluster 1 has 106 items.

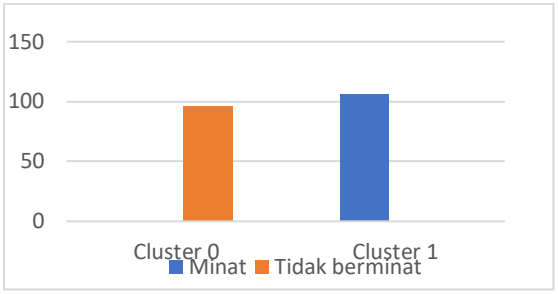

Fig. 7: based on interest chart

b. Grouping based on gender

After the results of the specialization cluster are completed, it can also be concluded that the ratio between female and male students in each cluster 0 and cluster 1 is quite significant. Out of 202 total student data, there are 137 female students and 65 male students. The percentage of the overall data will be calculated with the following formula

- In cluster 0, which totaled 96 items, there were 61 female students and 35 male students.
- While cluster 1 which amounted to 106 items, there were 76 female students and 30 male students.

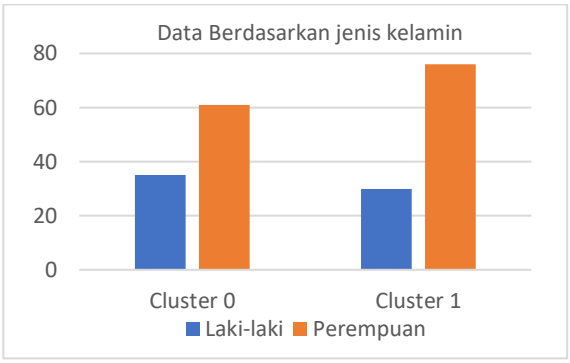

Fig 8: based on gender chart

c. Grouping based on their knowledge of the Informatics Engineering department.

Of the 202 total data, there are 155 students who know about informatics engineering majors and 47 students who did not know about informatics engineering majors before. After calculating based on the percentage calculation, the results are obtained as shown in the figure below.

Fig. 9: Knowlege Percentage chart

d. Grouping based on their liking for subjects related to Informatics.

Out of 202 students, it was found that 143 students liked subjects related to computers and technology, while 59 students disliked these subjects. After calculating based on the percentage calculation, the results are obtained as shown in the figure below.

Fig. 10: Liked Percentage chart

4. Conclusion

1) The clustering process using K-Means algorithm with Davies-Bouldin Index (DBI) evaluation shows that the best k value is k=2, with DBI of 0.527.
2) The clustering model was improved by adding some characteristics such as gender, knowledge of majors, and their liking for computer-related subjects.
3) The analysis showed that female students dominated both clusters and the clustering was effective enough to separate the students' interest data. The results can be used by schools to improve students' understanding of the major through guidance or recommendation systems.

5. Suggestions

1) Provide more access to computer labs, and innovative tools to support the learning of interested students.
2) More intensive socialization can increase their awareness and interest in this major in the future.
3) Future research can use other algorithms such as DBSCAN or Random Forest to compare with K-Means in improving the accuracy of student interest grouping.

References

[1] Arofah, S. N., & Marisa, F. (2018). Penerapan Data Mining untuk Mengetahui Minat Siswa pada Pelajaran Matematika menggunakan Metode K-Means Clustering. *JOINTECS (Journal of Information Technology and Computer Science)*, *3*(2), 85–90. https://doi.org/10.31328/jointecs.v3i2.787

[2] Hariyanto, D. C., Harini, S., & Chamidy, T. (2024). K-Means Clustering Dalam Pengelompokan Relevansi Pekerjaan S1 Informatika (Studi Kasus Jurusan Teknik Informatka Umm Malang). *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *9*(2), 782–797. https://doi.org/10.29100/jipi.v9i2.5507

[3] Mardika, P. D. (2023). Algoritma K-Means Untuk Mengetahui Minat Siswa Terhadap Jurusan Teknik Informatika. *Faktor Exacta*, *16*(2). https://doi.org/10.30998/faktorexacta.v16i2.17067

[4] Nurul Badriyah, Hozairi, & Miftahul Walid. (2023). Penentuan Bidang Minat Tugas Akhir Mahasiswa Teknik Informatika Universitas Islam Madura Menggunakan Metode K-Means. *Jurnal Informatika Teknologi Dan Sains (Jinteks)*, *5*(4), 566–572. https://doi.org/10.51401/jinteks.v5i4.2782

[5] Prihati, Y., Suwarno, & Dharmawan, A. (2019). Implementasi Algoritma K-Means Untuk Pemetaan Prestasi Akademik Siswa Disekolah Dasar Terang Bagi Bangsa Pati. *Kinabalu*, *11*(2), 50–57.

[6] Syahra, Y., Syahril, M., & Y, Y. (2019). Implementasi Data Mining Dengan Menggunakan Algoritma Fuzzy Subtractive Clustering Dalam Pengelompokan Nilai Untuk Menentukan Minat Belajar Siswa Smp Primbana Medan. *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika Dan Komputer)*, *17*(1), 54. https://doi.org/10.53513/jis.v17i1.113

[7] Widodo, W., & Wahyuni, D. (2017). Implementasi Algoritma K-Means Clustering Untuk Mengetahui Bidang Skripsi Mahasiswa Multimedia Pendidikan Teknik Informatika Dan Komputer Universitas Negeri Jakarta. *PINTER : Jurnal Pendidikan Teknik Informatika Dan Komputer*, *1*(2), 157–166. https://doi.org/10.21009/pinter.1.2.10

[8] Yuniarti, D. A. F., Kartika, D. L., & Prianggono, A. (2022). Analisis Minat Dan Motivasi Belajar Mahasiswa Teknik Informatika Pada Mata Kuliah Matematika. *JPMI (Jurnal Pendidikan Matematika Indonesia)*, *7*(1), 47. https://doi.org/10.26737/jpmi.v7i1.3437