Journal of Information Systems and Informatics
Vol. 5, No. 4, December 2023 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v5i4.576 Published By DRPM-UBD

Stream Clustering for Selection Recommendations Using K-Means Algorithm: A Case Study in the Informatics Study Program

Riska Fahmita Anggraini[1], Siti Sau'da[2]

[1,2]Informatics Departement, Bina Darma University, Palembang, Indonesia
Email: [1]fhmita10@gmail.com, [2]siti_sauda@binadarma.ac.id

Abstract

Concentration Stream for a major is a process where students focus their attention on a specific discipline according to their interests. The purpose of specialization is to better orient students to the knowledge they have gained from previous courses, so that they can have a clearer focus. In the Informatics Engineering study program at Bina Darma University there are 3 concentrations, namely: Software Engineering, Network Engineering, Data Analytics. The absence of a system that helps students choose a major concentration makes it quite difficult for students to know their academic abilities. By looking at these problems, this research aims to build a Recommendation system for selecting Mk-Stream Concentrations using the K-Means grouping approach using the K-Means cluster method. Where student academic achievement data from the first semester to the 4th semester is used as a variable in the calculations.

Keywords: Software, MK-Stream, K-Means, Clustering, Data Analytics

1. INTRODUCTION

The Informatics Engineering study program is one of the study programs available at universities which aims to produce graduates who are able to design, develop and implement information technology solutions. In the Informatics Engineering department at Bina Darma University, there is a concentration/specialization called Stream. In the study program in question, students will explore fundamental principles [1]. programming, data structures, algorithms, application development, computer networks, databases, information security, and web technology. Informatics Engineering students will also be trained to develop skills in problem analysis, critical thinking skills, and good communication skills [2]. Graduates of this study program have excellent career prospects in the field of information technology, ranging from being programmers, systems analysts, network administrators, web developers, to IT managers [3].

Choosing a concentration in student academic activities is not an easy thing because it really depends on the student's abilities, therefore careful consideration is needed so that students do not make a mistake in choosing the desired concentration. This often happens when final semester students do their final assignment but it does not match their field of ability [4]. Choosing a concentration haphazardly without careful consideration can have a negative impact on students, namely difficulty in absorbing lecture material. Therefore, a special method is needed that students can use to determine student concentration. One of the methods used is the K-Means method [5].

This K-means method is a method of clustering function so that data that has the same characteristics is grouped in the same cluster. Clustering is a technique used for data mining functionality. The clustering algorithm is grouping data into certain data groups (clusters) [6]. The main objective of this research is to implement data mining techniques using the K-Means algorithm to group data from students taking the Informatics Engineering-S1 study program based on their academic performance. It is hoped that the results of this grouping will provide new and useful information as advice in determining the right concentration path for students [7]. The data analyzed includes information about students, their academic records, such as KHS and transcripts from semesters 1 to 4. The information resulting from this analysis can be one of the options considered by study programs to provide advice to students regarding the choice of appropriate concentration paths. with their abilities [8].

Concentration or specialization refers to the focus that students have on a particular field of study that suits their interests . The aim of this concentration is to direct students more deeply into the knowledge they have acquired from previous courses. Therefore, this research aims to apply data mining methods, especially grouping techniques using the K-Means algorithm [9]. This is done so that students can be grouped based on their academic abilities, and the results of this grouping can be an alternative for study programs in providing recommendations to students about the appropriate concentration path for them to choose [10].

2. METHODS

K-Means algorithm is a popular clustering technique used in data mining and machine learning. It is a partitioning method that aims to divide a dataset into K distinct, non-overlapping clusters. The algorithm works by iteratively assigning data points to the nearest cluster center and then updating the cluster centers based on the mean of the data points in each cluster. This process continues until convergence, resulting in K clusters where each data point belongs to the cluster with the nearest mean. In summary, the K-Means algorithm is used for unsupervised clustering and helps group similar data points together into

clusters, making it easier to analyze and understand the underlying structure in the data. To provide guidance in designing this research, a structured plan is needed that outlines the steps. This plan is a series of stages carried out to solve the problems in this research. This research has a series of processes which will be shown in the illustration Figure 1.

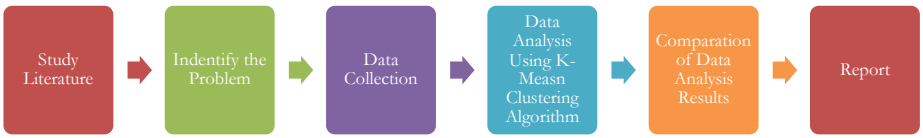

Figure 1. Clustering Algorithm K-Means Methods

Based on the structure of the research framework described previously, the discussion steps in the research are as follows:

2.1. Study Literature

In this step, researchers use various available document sources to access data and information that is relevant for research. Researchers extract data from various sources, including books, journals and websites that are relevant to the problem being investigated. The main goal is to collect information that will be useful in the research process [12].

2.2. Identify the Problem

In the first stage, researchers identify the problem. The purpose of this step is to help researchers collect information about the problems contained in the research object, with the aim of finding the root of the problem which will become the basis of this research [11].

2.3. Data Collection

In this stage, researchers apply data collection methods, which involve field observations and document analysis. Field observations were carried out directly to collect relevant information [13].

3. RESULTS AND DISCUSSION

A result was found where the recommendations for stream selection for informatics study program students using the k-means algorithm clustering method would be selected automatically when calculating variable variables to

better direct students to choose a concentration according to what they desired and the student's abilities.

In the first stage, the researcher identifies the problem. The purpose of this step is to assist researchers in identifying what problems occur, for example, in this study regarding the K-Means algorithm clustering method in student stream selection recommendations.

3.1 Manual K-Means Clustering Calculation

The application of the K-Means method for grouping was carried out on 172 examples of student data. The data generated from the following process can produce the required group data, which can later be applied in the classification stage. To identify recommended data groups for selecting majors in higher education, the clustering method is used. One of the techniques used is the K-Means method, which is a distance-based clustering algorithm to group data into several groups. It is important to note that this method only applies to attributes that have numeric values The stages in computing K-Means Clustering include: Identifying the desired number of groups for data, Selecting the center point for each group, Adding up the distance between each data point at the center point of each cluster, Grouping objects into clusters based on proximity to each other. center point (centroid) and If the results of the new data grouping are identical to the results of the previous grouping, then the calculation is considered complete.

3.1.1 Data Collection

Data analysis is needed to identify students who have potential, applying the K-Means clustering approach to group them according to their abilities and expertise. This academic data includes student grades derived from the results of their studies each semester. To carry out analysis and grouping potential students, we need sample data. This sample data was obtained by accessing student achievement data for the 2020 Informatics Engineering Study Program from student data on the server. We have taken care of the necessary permissions from DSTI to access this data. The object of our research is the academic data of students from the 2020 and 2021 classes, as in the Figure below.

Table 1. Student Academic Data 2020-2021

Nim	Course Name	sks	Score	semester
201410026	Multimedia	2	E	2
201420001	Work lectures	2	E	0
201420001	Algorithms and Programming	3	A	1
201420001	Computer architecture and organization	3	B	1

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Nim	Course Name	sks	Score	semester
201420001	Calculus	3	B	1
201420001	Introduction to Information Technology	3	A	1
201420001	Programming practicum	3	A	1
201420001	Computer Network 1	3	B	2
201420001	Discrete Mathematics	3	B	2
201420001	Introduction to Multimedia	3	A	2
201420001	Data Structure Procedures and Advanced Algorithms	2	A	3
201420001	Operating System (OS)	2	A	3
201420001	Advanced data structures and Algorithms	2	C	3
201420001	Numerical analysis	4	A	3
201420001	Database	2	A	3
201420001	Human and computer interaction	2	A	3
201420001	Computer Network 2	2	A	4
201420001	Database practicum	4	A	4
201420001	Software engineering	2	A	4
201420001	Probalitas Statistics	4	A	4
201420001	Web programing	2	A	4
201420001	Linear algebra	2	A	4
201420001	Non rational database	4	A	4
201420001	Computer graphics	2	A	4
201420001	IT enterpreneurship	2	A	4
201420001	Artificial Intelligence	2	A	4
201420001	Industry Introduction Lecture	2	A	4
201420001	Object-oriented programming	4	B	4
201420001	Graph theory	2	A	4

The data above is a display of the entire data from the 2020-2021 class of students after being combined into one CSV file. Scraping was carried out for each category, each of which took 100 pages. The total amount of data after being combined was 1,609 rows and 6 columns.

3.2 Data preprocessing

3.2.1 Data Cleaning

In the data cleaning stage regarding student information, KHS, and transcripts, a number of actions will be carried out. The initial step is to delete empty data (null) and unrelated (irrelevant) data, inconsistent data, and data entered with errors. After that, in the data merging process, we will create a new data set using

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

the attributes that have been previously defined. The attributes that will be included in this new dataset include Student Identification Number, GPA, Semester Achievement Index 1 to 4, average value of supporting courses in the fields of SC (Computer Systems) and RPLD (Software and Data Engineering), as well as average -average course grades in the SC and RPLD specialization fields, as well as other specializations.

Table 2. Cleaned Data

Nim	Course Name	sks	Score	Semester
201420001	Algorithms and Programming	3	A	1
201420001	Computer architecture and organization	3	B	1
201420001	Programming practicum	3	A	1
201420001	Computer Network 1	3	B	2
201420001	Database	2	A	3
201420001	Computer Network 2	4	A	3
201420001	Database Practicum	2	A	3
201420001	Web programaming	4	B	3
201420001	Non rational database	4	A	4
201420002	Algorithms and Programming	3	E	1
201420002	Computer architecture and organization	3	E	1
201420002	Programming practicum	3	E	1
201420003	Computer Network 1	3	E	1
201420003	Database	3	A	3
201420003	Computer Network 2	3	B	3
201420003	Database Practicum	3	A	1
201420003	Web programaming	3	B	2
201420003	Non rational database	3	A	3
201420004	Algorithms and Programming	3	B	3
201420004	Computer architecture and organization	2	A	3
201420004	Computer architecture and organization	4	A	4
201420004	Programming practicum	3	A	1
201420004	Computer Network 1	3	C	1
201420004	Database	3	A	1
201420004	Computer Network 2	3	B	1

3.2.1 Data representation

In the Informatics Engineering study program at Bina Darma University there are 3 concentrations, namely: Software Engineering, Network Engineering, Data

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Analytics. The following are several Stream courses in the informatics engineering study program.

Tabel 3. Initial Center Point (Centroid) Determination Table

Nama Matakuliah	A	B	C	D(P1,C)	D(P1,1)
Algoritma & Pemrograman	3	B	1	0	1
Arisitektur dan organisasi Komputer	3	B	2	1	0
Praktikum pemrograman	2	A	3	2,23	1,41
Jaringan komputer 1	3	A	1	0	1
Basis data	4	B	3	2,33	1,41
Jaringan komputer 2	2	A	3	2,33	1,41
Praktikum basis data	4	B	3	2,33	1,41
Web programing	3	C	1	0	1
Basis data non rekasional	2	A	3	2,23	1,41

Tabel 2. Tabel Centroid

Centroid	**A**	**B**
1	3	1
2	3	2

Measure the distance of data to the nearest centroid using the Euclidean Distance method by calculating it as follows.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Where Xi is the ith object in cluster c, Yi is the ith object in cluster y, and N is the total number of objects in the cluster. For example, let's calculate the distance between object 1 and the center point in cluster 1.

$$d(p_1, c_1)=\sqrt{(p_1a\text{-}c_1a).^2+(p_1b\text{-}c_1b).^2}$$

$$\sqrt{(3\text{-}3).^2+(A\text{-}A).^2+(1\text{-}1).^2}$$

The total distance of the 2nd object to the center of cluster 2, namely:

$$d(p_2, c_1)=\sqrt{(3\text{-}3).^2+(2\text{-}1).^2}$$

The total distance of the 3rd object to the center of cluster 3, namely:

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

$$d(p_3, c_1)=\sqrt{(2-3).^2+(3-1).^2}=2{,}23$$

The total distance of the 4th object to the center of cluster 4, namely:

$$d(p_4, c_1)=\sqrt{(3-3).^2+(1-1).^2}=0$$

The total distance of the 5th object to the center of cluster 5, namely:

$$d(p_5, c_1)=\sqrt{(4-3).^2+(3-1).^2}=2{,}23$$

The total distance of object 6 to the center of cluster 6, namely:

$$d(p_6, c_1)=\sqrt{(2-3).^2+(3-1).^2}=2{,}23$$

The total distance of the 7th object to the center of cluster 7, namely:

$$d(p_7, c_1)=\sqrt{(4-3).^2+(3-1).^2}=2{,}23$$

The total distance of the 8th object to the center of cluster 8, namely:

$$d(p_8, c_1)=\sqrt{(3-3).^2+(1-1).^2}=0$$

The total distance of object 9 to the center of cluster 9, namely:

$$d(p_9, c_1)=\sqrt{(4-3).^2+(4-1).^2}=3{,}16$$

The total distance of object 10 to the center of cluster 10, namely:

$$d(p_{10}, c_1)=\sqrt{(2-3).^2+(3-1).^2}=2{,}23$$

The total distance of object 11 to the center of cluster 11, namely:

$$d(p_{11}, c_2) = \sqrt{(p_1 a - c_2 a).^2+(p_1 b - c_2 b).^2}$$
$$=\sqrt{(3-3).^2+(1-2).^2}=1$$

The total distance of object 12 to the center of cluster 12, namely:

$$d(p_{12}, c_2)=\sqrt{(3-3).^2+(2-2).^2}=0$$

The total distance of object 13 to the center of cluster 13, namely:

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

$$d(p_{13}, c_2)=\sqrt{(2-3).^2+(3-2).^2}=1{,}41$$

The total distance of object 14 to the center of cluster 14, namely:

$$d(p_{14}, c_2)=\sqrt{(3-3).^2+(1-2).^2}=1$$

The total distance of object 15 to the center of cluster 15, namely:

$$d(p_{15}, c_2)=\sqrt{(4-3).^2+(3-2).^2}=1{,}41$$

The total distance of object 16 to the center of cluster 16, namely:

$$d(p_{16}, c_2)=\sqrt{(2-3).^2+(3-2).^2}=1{,}41$$

The total distance of object 17 to the center of cluster 17, namely:

$$d\ (p_{17}, c_2)=\sqrt{(4-3).^2+(3-2).^2}=1{,}41$$

The total distance of the 18th object to the center of cluster 18, namely:

$$d(p_{18}, c_2)=\sqrt{(3-3).^2+(1-2).^2}=1$$

The total distance of object 19 to the center of cluster 19, namely:

$$d(p_{19}, c_2)=\sqrt{(4-3).^2+(4-2).^2}=2{,}23$$

The total distance of 20 objects to the center of cluster 20, namely:

$$d(p_{20}, c_2)=\sqrt{(2-3).^2+(3-2).^2}=1{,}41$$

Centroid Point,

$$X_{baru} = \frac{x_{1+}x_4 + x_8}{3} = \frac{3+3+3}{3} = 3$$

$$Y_{baru} = \frac{1+1+1}{3} = 1$$

$$Cluster = \frac{x_2 + x_3 + x_5 + x_6 + x_7 + x_9 + x_{10}}{3}$$

$$= \frac{3+2+4+2+4+4+2}{3} = \frac{21}{3} = 3$$

$$Y_{baru} = \frac{2+3+3+3+3+4+3}{7} = \frac{21}{7} = 3$$

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.2.2 Model Evaluation

Model evaluation in the research "Stream Clustering for Selection Recommendations Using the K-Means Algorithm: A Case Study in the Informatics Study Program" is crucial to measure the quality of clustering and the model's performance in dealing with streaming data. Metrics such as Inertia and Silhouette Score are used to assess the quality of clustering, while metrics like Modified Rand Index and Dynamic Euclidean Distance are employed to evaluate the model's performance on the evolving stream of data. Furthermore, real-time evaluation by periodically calculating the Silhouette Score can provide insights into the model's performance in a continuous data stream scenario. Cross-validation helps measure the overall model performance, while business impact evaluation is a crucial step to understand how much value the model adds in delivering better selection recommendations in the Informatics study program. The combination of these metrics and evaluations helps gauge the success and relevance of the model in the context of the conducted research and enables necessary improvements. At this stage, several plots will be displayed to see the performance results of the K-Means cluster method. The image below is a PCA plot which shows a two-dimensional visualization depicting data that has been dimensionally reduced using PCA.

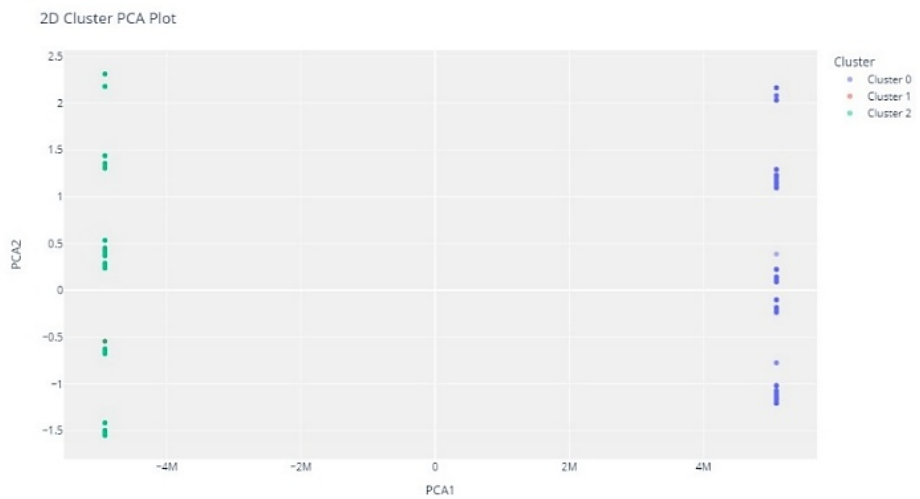

Figure 2. PCA Plot K-Means

Then the image below is a Distribution Plot. This plot displays the distribution of data or how much data each cluster has. It can be seen that cluster 1 is the cluster that has the most data with more than 800 data. Apart from that, there is also other data which is explained in the K-Means Plot Distribution image. In cluster 0 there are 380 data and in cluster 2 there are 450 data.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

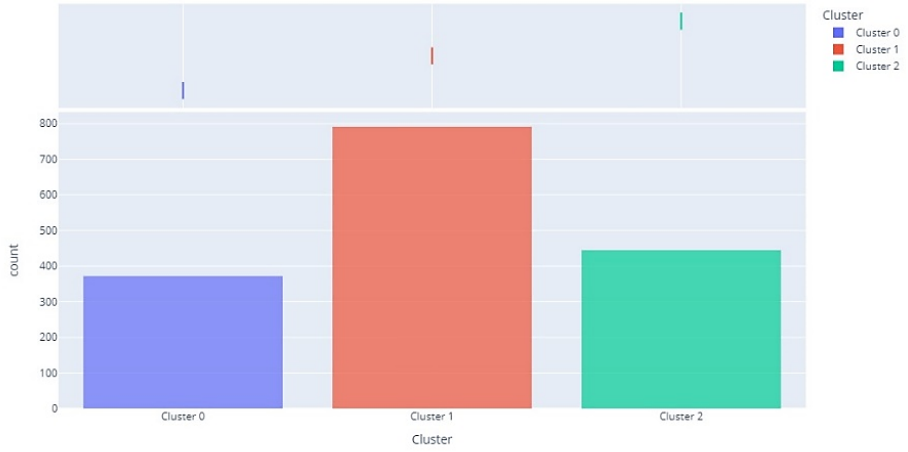

Figure 3. K-Means Plot Distribution

And the following image is a Distance Plot. This plot displays the distance from the point or center of the cluster that has been created. In this plot, it can be seen that K-Means can separate clusters quite well in clusters 1 and 3.

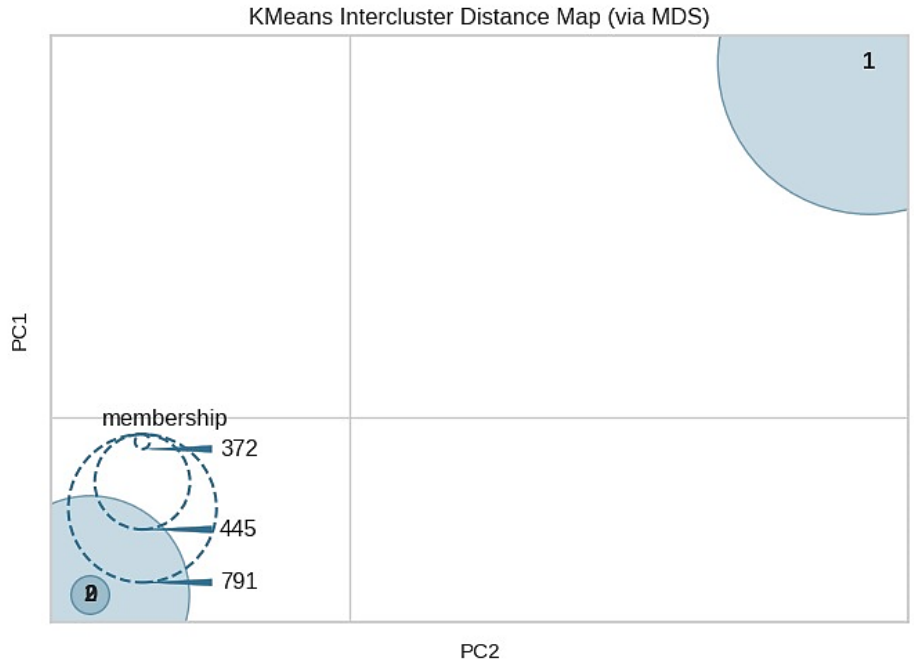

Figure 4. Distance Plot K-Means

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.3 Determining the Path of Interest

The characteristics of the three clusters formed will be studied further in order to gain insight into the extent to which they are compatible with existing vocational pathways. This suitability assessment will be carried out by connecting it using the attributes in the specialization, the average course value (MK) in the SC specialization, and the average MK value in the RPLD specialization. In this way, we can explore information and patterns of student characteristics based on their academic abilities in accordance with existing specialization paths. In this cluster, 87% of the total 1069 students chose the Software Engineering route, while 75% of the 1069 students chose the Network route. This indicates that this cluster is dominated by students who choose Software Engineering rather than Networking. In addition, 97.5% of 1069 students chose the Data Analytics pathway.

3.4 Measuring the Level of Recommendation Accuracy

To assess the level of accuracy of the recommendations produced, a comparison is carried out with the previous recommendation system to evaluate the accuracy regarding the specialization path that the student will choose. Which resulted in a comparison using 1069 data showing that the accuracy level of the recommendations made was 81%, while the accuracy level of the old recommendation system was only 7.55%.

3.4 Deployment

In this step, create a report based on the insights obtained through the data mining process. This was achieved by applying a data mining method that uses the K-Means algorithm on student data registered in the Undergraduate Informatics Engineering Study Program in the 2020-2021 academic year, successfully providing suggestions for selecting a specialization path that is in accordance with the student's academic competence. So this can be used as an alternative recommendation provided by the Undergraduate Informatics Engineering study program to students. This information can be a guide for students who feel confused when choosing a stream that suits their level of academic expertise.

4. CONCLUSION

From this research, the use of the grouping method using the K-Means method in data analysis resulted in the formation of three groups of students based on their academic abilities. It is hoped that this analysis will have a positive impact on IT study programs in the future, as one of the many recommendation options that can be suggested to students in the context of determining a suitable

specialization path based on their level of academic strength. The results of the comparison used to evaluate the level of accuracy of recommendations show that the level of accuracy of recommendations based on the specialization chosen by students reaches 81%. Meanwhile, the level of accuracy of the system used to provide previous advice related to determining the path of interest taken by students only reached 7.55%.

REFERENCES

[1] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *J. Tekno Kompak*, vol. 15, no. 2, p. 25, 2021, doi: 10.33365/jtk.v15i2.1162.

[2] R. Adrianto and A. Fahmi, "Penerapan Metode Clustering Dengan Algoritma K-Means Untuk Rekomendasi Pemilihan Jalur Peminatan Sesuai Kemampuan Pada Progam Studi Teknik Informatika - S1 Universitas Dian Nuswantoro," *JOINS (Journal Inf. Syst.*, vol. 1, no. 2, pp. 101–116, 2019, [Online]. Available: http://publikasi.dinus.ac.id/index.php/joins/article/view/1302

[3] Fina Nasari and S. Surya Darma, "Penerapan K-Means Clustering Pada Data Penerimaan Mahasiswa Baru," *Semin. Nas. Teknol. Inf. dan Multimed. 2015*, pp. 73–78, 2015.

[4] M. ISTONINGTYAS, "Penentuan Jurusan ke Perguruan Tinggi Menggunakan Metode Clustering di SMAN 3 Kuala Tungkal," *J. Process.*, vol. 13, no. 2, 2018, [Online]. Available: http://ejournal.stikom-db.ac.id/index.php/processor/article/view/352

[5] F. Yunita, "Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru," *Sistemasi*, vol. 7, no. 3, p. 238, 2018, doi: 10.32520/stmsi.v7i3.388.

[6] Y. R. Sari, A. Sudewa, D. A. Lestari, and T. I. Jaya, "Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 2, p. 192, 2020, doi: 10.24114/cess.v5i2.18519.

[7] K. Handoko, "Penerapan Data Mining Dalam Meningkatkan Mutu Pembelajaran Pada Instansi Perguruan Tinggi Menggunakan Metode K-Means Clustering (Studi Kasus Di Program Studi Tkj Akademi Komunitas Solok Selatan)," *J. Teknol. dan Sist. Inf.*, vol. 02, no. 03, pp. 31–40, 2016, [Online]. Available: http://teknosi.fti.unand.id/index.php/teknosi/article/view/70

[8] A. Situmorang, A. Arifin, I. Rusilpan, and C. Juliane, "Analisa dan Penerapan Metode Algoritma K-Means Clustering Untuk Mengidentifikasi Rekomendasi Kategori Baru Pada List Movie IMDb," *J. Media Inform. Budidarma*, vol. 6, no. 4, p. 2171, 2022, doi: 10.30865/mib.v6i4.4729.

[9] I. Mahmud, A. D. Indriyanti, and I. Lazulfa, "Penerapan Algoritma K-Means Clustering Sebagai Strategi Promosi Penerimaan Mahasiswa Baru

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Pada Universitas Hasyim Asy'ari Jombang," *Inovate*, vol. 4, no. 2, pp. 20–27, 2020.

[10] M. Mustofa, "Penerapan Algoritma K-Means Clustering pada Karakter Permainan Multiplayer Online Battle Arena," *J. Inform.*, vol. 6, no. 2, pp. 246–254, 2019, doi: 10.31311/ji.v6i2.6096.

[11] M. Benri, H. Metisen, and S. Latipa, "Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila," *J. Media Infotama*, vol. 11, no. 2, pp. 110–118, 2015, [Online]. Available: https://core.ac.uk/download/pdf/287160954.pdf

[12] M. L. Sibuea and A. Safta, "Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustring," *Jurteksi*, vol. 4, no. 1, pp. 85–92, 2017, doi: 10.33330/jurteksi.v4i1.28.

[13] J. Hutagalung and F. Sonata, "Penerapan Metode K-Means Untuk Menganalisis Minat Nasabah," *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 1187, 2021, doi: 10.30865/mib.v5i3.3113.