http://ejournal.seaninstitute.or.id/index.php/InfoSains
Jurnal Info Sains : Informatika dan Sains, Volume 13, No 02, 2023
E-ISSN.2797-7889, P-ISSN.2089-3329

Alumni Data Grouping Using the K-Means Clustering Method for Study Program Curriculum Development

[1] Penda Sudarto Hasugian, [2] Jijon Raphita Sagala, [3] Lela Dwi Ani
[1,2,3] Informatics Engineering, STMIK Pelita Nusantara
E-mail: penda.hasugian@gmail.com[1],sisagala@gmail.com[2], lelad@gmail.com[3]

Keywords

Datamining, K-means method, Data grouping, Rapid Miner

Abstract. Application of Datamining by applying the k-means clustering method to classify STMIK Pelita Nusantara alumni data as a basis for developing study program curricula that are more relevant to the needs of the world of work or industrial needs. Where the K-Means Clustering Method is used to group alumni based on similar characteristics they have, such as personal data, academic achievement, areas of expertise, and job information after graduating from college. The research data source used is graduate data for the 2021/2022 academic year. The data collection method was carried out by distributing questionnaires directly to alumni. The application of the k-means method is carried out by forming 2 groups (clusters), namely C1 = Liner and C2 = Not Linear. Data testing is also carried out using the rapid miner application. So that by grouping alumni data, it is hoped that tertiary institutions can identify the needs and preferences of alumni for the study programs followed so that they can develop study program curricula that are more targeted and in accordance with the needs of the job market.

1. INTRODUCTION

STMIK Pelita Nusantara is one of the tertiary institutions in North Sumatra, precisely in the city of Medan which has 6 study programs ranging from diploma three (D-III) to undergraduate level (S1) in Computer Science and already has many alumni. Where universities are responsible for creating quality graduates who suit the needs of the world of work. This is important for tertiary institutions to develop study program curricula that are right on target and in accordance with the needs of the world of work. However, in developing study program curricula, there is often a gap between the needs of the world of work and the curriculum implemented in tertiary institutions. To overcome this problem, accurate information is needed about the characteristics of alumni and the jobs they undertake after graduating.[1]. This alumni data can be used to develop study program curricula that are more suited to the needs of the world of work. However, the available alumni data is usually raw data which is difficult to process and analyze. Therefore, this research will carry out a data mining process[2]. Data mining is a process of extracting or collecting important information from a cluster or large data set, this data mining process applies the k-means clustering method to classify alumni data based on their similar characteristics[3]. This grouping of alumni data will help tertiary institutions to identify the needs and linkages of alumni to the chosen study program. Thus, tertiary institutions can develop study program curricula that are more targeted and in accordance with the needs of industry or the world of work[4].

2. METHODS

Datamining

Data mining is a process of extracting data or collecting important information from a cluster or large data collection. The data mining process often uses statistical methods, mathematics, to utilize data processing technology or applications such as rapid miner[5]. In the KDD (knowledge discovery in databases) process, many concepts and techniques are used in the data mining process. The process requires several steps to get the desired data[6]. The stages in the KDD process are as follows:

A. Data Selection
   The data used is data from STMIK Pelita Nusantara graduates taken in the 2021/2022 academic year.
B. Data Cleaning

Data cleaning is carried out to clean incomplete data such as missing data, invalid data and typos. because incomplete data can reduce the quality and accuracy of data mining results. Data cleaning also has an impact on the performance of the data mining system because it reduces the amount and complexity of data processed[7].

C. Data integration
Data integration is carried out on attributes that identify the unique entity STMIK Pelita Nusantara, data integration attributes must be carried out carefully, because data integration errors can cause biased data results and can mislead the actions taken later.

D. Data transformation
Is a transformation process on selected data, so that the data is suitable for the data mining process.[8]

E. Data Mining
Data mining is the process of finding and analyzing large amounts of data with the aim of finding interesting patterns or information from large amounts of stored data using certain techniques or methods. The appropriate technique, method, or algorithm really depends on the objectives and process of knowledge discovery in databases (KDD) as a whole[9]. This stage is the core of the knowledge discovery in databases (KDD) stage which is carried out to analyze data that has been cleaned.

F. Pattern evaluation
In this step, the results of data mining techniques in the form of patterns and models are evaluated to determine whether they can actually be achieved.

G. Knowledge
The last step in the data mining process is to form a decision or action based on the results of the analysis.

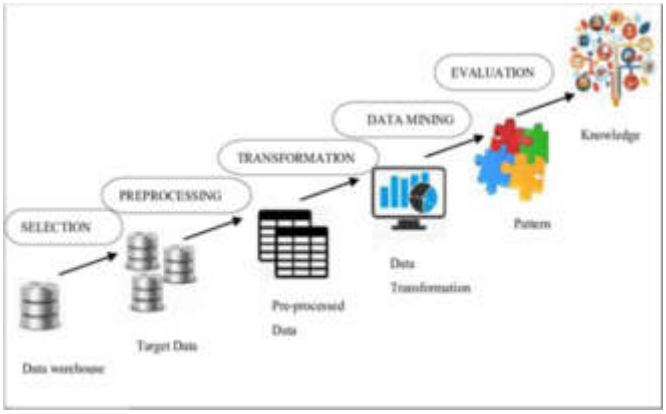


Figure 1. Data mining stages

K-Means Algorithm

K-Means is an iterative clustering algorithm. This method partitions the data into clusters/groups so that data that has the same characteristics are grouped into the same cluster and data that has different characteristics are grouped into another group.[10]. The K-Means algorithm basically performs two processes, namely the process of detecting the location of the center of each cluster and the process of searching for members from each cluster. Stages of implementing the K-Means algorithm[11]:

a. Determine k as the number of clusters you want to form.
b. Generate the initial k centroids (cluster center points) randomly.
c. Calculate the distance of each data to each centroid.
d. Each data chooses the nearest centroid.
e. Determine the new centroid position by calculating the average value of the data located at the same centroid.
f. Return to step 3 if the position of the new centroid and the old centroid are not the same

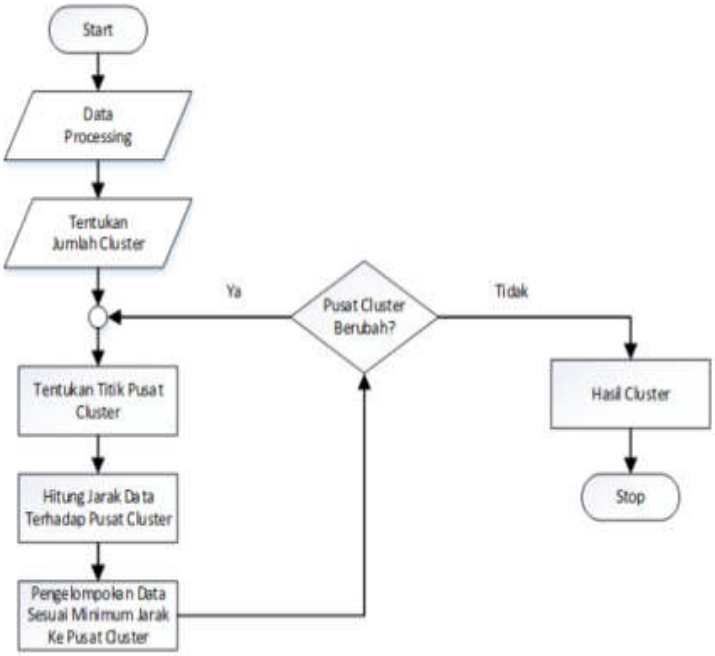


Figure 2. Stages of the K-Means Clustering Method

Research Stages

The research stages carried out are as shown in Figure 3:

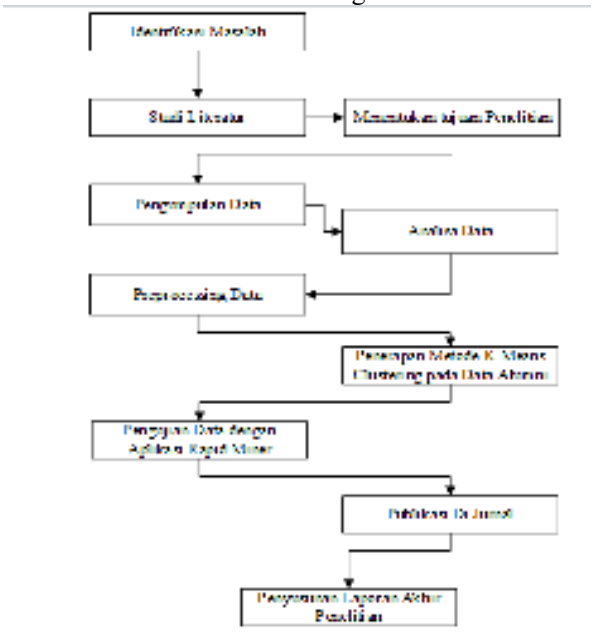


Figure 3. Research Stages

3. RESULTS AND DISCUSSION

Data analysis

The data used comes from STMIK Pelita Nusantara alumni data. The data will be used in the calculation process with the k-means algorithm as shown in table 1 below.

Table 1. Alumni Data

No	Nama	Jurusan	Tahun Lulus	Bidang Keahlian	Pekerjaan	Tempat Pekerjaan
1	Alumni001	Teknik InformatiKa	2022	Bahasa Pemrograman PHP	Web Developer	RS. Grandmed
2	Alumni002	Manajemen Informatika	2022	Bahasa PemrogramanVisual	Staf Administrasi	CV. Oloan Komuter
3	Alumni003	Teknik InformatiKa	2022	Bahasa PemrogramanVisual	Mekanik	Bengkel Honda
4	Alumni004	Manajemen Informatika	2022	Dara Mining	Marketing	PT. Mega Auto Finance
5	Alumni005	Teknik InformatiKa	2022	Bahasa PemrogramanVisual	Programmer	Digitalindo
6	Alumni006	Teknik InformatiKa	2022	Bahasa Pemrograman PHP	Programmer	PT. Media Creative Nusantara
7	Alumni007	Teknik InformatiKa	2022	Bahasa Inggris	Karyawan Swalayan	Matahari
8	Alumni008	Manajemen Informatika	2022	Bahasa Pemroraman Java	Programmer	SMP Wira Nusantara
9	Alumni009	Manajemen Informatika	2022	Dara Mining	Staf Administrasi	PT. Industri Medan Deli
10	Alumni010	Teknik InformatiKa	2022	Bahasa Pemroraman Java	Network Developer	CV. Makmur Network
11	Alumni011	Manajemen Informatika	2022	Bahasa Pemrograman PHP	Web Developer	Percetakan
12	Alumni012	Teknik InformatiKa	2022	Bahasa Inggris	Staf Produksi	PT. Sagami

From the alumni data obtained, 12 data samples were taken for the calculation process using the k-means clustering method. To make it easier to apply the k-means clustering method, the data transformation process was carried out as in table 2 below:

Table 2. Data transformation

No	Name	Major	Skill	Work
1	Alumni001	1	1	1
2	Alumni002	2	2	5
3	Alumni003	1	2	7
4	Alumni004	2	5	8
5	Alumni 005	1	2	2
6	Alumni006	1	1	2
7	Alumni007	1	4	6
8	Alumni008	2	3	2
9	Alumni009	2	5	5
10	Alumni010	1	3	3
11	Alumni011	2	1	1
12	Alumni012	1	4	9

Application of the K-means Clustering Method

The application of the K-means method aims to group data according to the groups that will be formed[4]. The following are the stages in the calculation process:

1) Determine the number of data groups or number of Clusters
   The number of clusters formed is 2 clusters, so the value of k=2. Where C1 = Linear, C2 = Not Linear.
2) Defines the cluster center point
   The cluster center point is taken from the data that has been randomly transformed. The data used as the initial cluster center are the 1st alumni data and the 8th data.

Table. 3 Initial Centroid Centers

Clusters	Major	Skill	Work
C1	1	1	1
C2	2	3	2

3) Calculate the distance from each data to the center of the cluster using the formula:

$$d_{ij} = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{kj})^2}$$

Where is the distance of the first data at the center of the first cluster:

$d_{11} = \sqrt{(1-1)^2 + (1-1)^2 + (1-1)^2} = 0.000$
The first data distance to the center of the second cluster is:
$d_{12} = \sqrt{(1-2)^2 + (1-3)^2 + (1-2)^2} = 2.449$

The results of calculating the distance for each data to each Cluster center are shown in table 4.

Table.4 Calculation results of data for each Cluster center

No	Nama	Jurusan	Keahlian	Pekerjaan	C1	C2	Jarak Terdekat	Keterangan
1	Alumni001	1	1	1	0,000	2,449	0,000	Cluster 1
2	Alumni002	2	2	5	4,243	3,162	3,162	Cluster 2
3	Alumni003	1	2	7	6,083	5,196	5,196	Cluster 2
4	Alumni004	2	5	8	8,124	6,325	6,325	Cluster 2
5	Alumni005	1	2	2	1,414	1,414	1,414	Cluster 2
6	Alumni006	1	1	2	1,000	2,236	1,000	Cluster 1
7	Alumni007	1	4	6	5,831	4,243	4,243	Cluster 2
8	Alumni008	2	3	2	2,449	0,000	0,000	Cluster 2
9	Alumni009	2	5	5	5,745	3,606	3,606	Cluster 2
10	Alumni010	1	3	3	2,828	1,414	1,414	Cluster 2
11	Alumni011	2	1	1	1,000	2,236	1,000	Cluster 1
12	Alumni012	1	4	9	8,544	7,141	7,141	Cluster 2

From the results of the distance calculation, the data groups for each cluster are taken based on the closest or smallest distance, namely: C1 = data 1, 6 and 11. C2 = 2, 3, 4, 5, 7, 8, 9, 10 and 12.

4) Compute New Cluster Center

To calculate the new cluster center where the data used is data from each cluster. By using the formula:

$$c_{kj} = \frac{\sum_{h=1}^{p} y_{hj}}{p}; y_{hj} = x_{hj} \in cluster ke - k$$

For C1 there are 3 data, namely data 1,6 and 11. Where:

$$C_{11} = \frac{1+1+2}{3} = 1.333$$
$$C_{12} = \frac{1+1+1}{3} = 1.000$$
$$C_{13} = \frac{1+2+1}{3} = 1.333$$

For the 2nd Cluster (C2) there are 9 data, namely data to data 2, 3, 4, 5, 7, 8, 9, 10, and 12 where:

$$C_{21} = \frac{2+1+2+1+1+2+2+1+1}{9} = 1.444$$
$$C_{22} = \frac{2+2+5+2+4+3+5+3+4}{9} = 3.333$$
$$C_{23} = \frac{5+7+8+2+6+2+5+3+9}{9} = 5.222$$

So that the new cluster center obtained is as in table 5.

Table. 5 New Cluster Centers

Clusters	Major	Skill	Work
C1	1,333	1,000	1,333
C2	1,444	3,333	5,222

5) Recalculating the data distance to each new cluster center, until the data for each cluster remains constant. As in Table 6 and Table 7 the results of the 2nd iteration and 3rd iteration.

Table.6 Results of 2nd Iteration

No	Nama	Jurusan	Keahlian	Pekerjaan	C1	C2	Jarak Terdekat	Keterangan
1	Alumni001	1	1	1	0,938	5,766	0,938	Cluster 1
2	Alumni002	2	2	5	3,476	2,025	2,025	Cluster 2
3	Alumni003	1	2	7	5,430	1,841	1,841	Cluster 2
4	Alumni004	2	5	8	7,272	2,412	2,412	Cluster 2
5	Alumni005	1	2	2	0,693	4,452	0,693	Cluster 1
6	Alumni006	1	1	2	0,825	4,895	0,825	Cluster 1
7	Alumni007	1	4	6	5,028	0,623	0,623	Cluster 2
8	Alumni008	2	3	2	1,575	4,221	1,575	Cluster 1
9	Alumni009	2	5	5	4,846	1,917	1,917	Cluster 2
10	Alumni010	1	3	3	2,020	3,223	2,020	Cluster 1
11	Alumni011	2	1	1	1,039	5,778	1,039	Cluster 1
12	Alumni012	1	4	9	7,790	2,921	2,921	Cluster 2

Table.7 3rd Iteration Results

No	Nama	Jurusan	Keahlian	Pekerjaan	C1	C2	Jarak Terdekat	Keterangan
1	Alumni001	1	1	1	1,225	6,283	1,225	Cluster 1
2	Alumni002	2	2	5	3,240	2,025	2,025	Cluster 2
3	Alumni003	1	2	7	5,180	1,841	1,841	Cluster 2
4	Alumni004	2	5	8	6,964	2,412	2,412	Cluster 2
5	Alumni005	1	2	2	0,408	4,452	0,408	Cluster 1
6	Alumni006	1	1	2	0,913	4,895	0,913	Cluster 1
7	Alumni007	1	4	6	4,708	0,623	0,623	Cluster 2
8	Alumni008	2	3	2	1,354	4,221	1,354	Cluster 1
9	Alumni009	2	5	5	4,528	1,917	1,917	Cluster 2
10	Alumni010	1	3	3	1,683	3,223	1,683	Cluster 1
11	Alumni011	2	1	1	1,354	5,778	1,354	Cluster 1
12	Alumni012	1	4	9	7,494	2,921	2,921	Cluster 2

In the 3rd iteration, the data position for each cluster is the same as the data position in the 2nd iteration. So the distance calculation process is stopped. Then the final results for each cluster C1 and C2 formed are:

1. Cluster 1 (C1) with cluster centers (1,333. 1,833 and 1,833) which are defined as Linear clusters namely alumni001, alumni005, alumni006, alumni008, alumni010, alumni011,
2. Cluster 2 (C2) with cluster centers (1,500, 3,667 and 6,667) which is a non-linear cluster, where there are 6 data included in this cluster namely alumni002, alumni003, alumni004, alumni007, alumni009, alumni012.

Testing with the Rapid Miner Application

Tests are carried out to adjust data calculations by applying the stages and the k-measn formula, the following stages of testing are carried out:

A. Data processing

The data to be tested is entered into rapidminer with the Excel data format as shown in Figure 4 below:

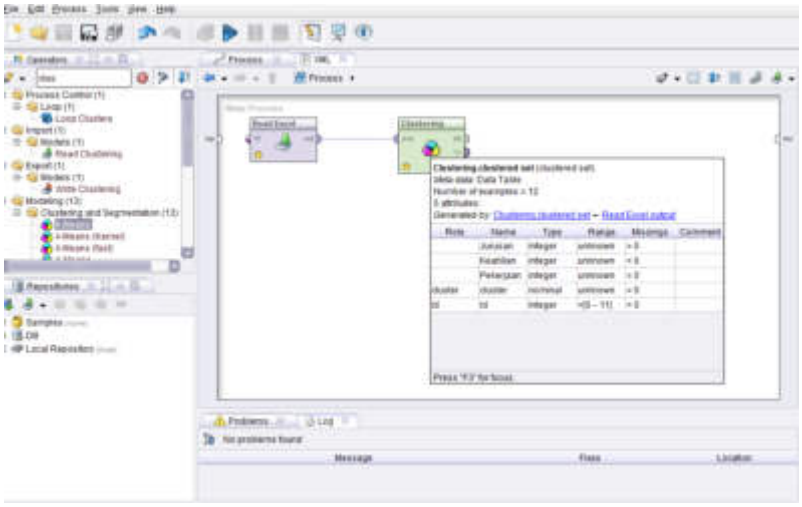
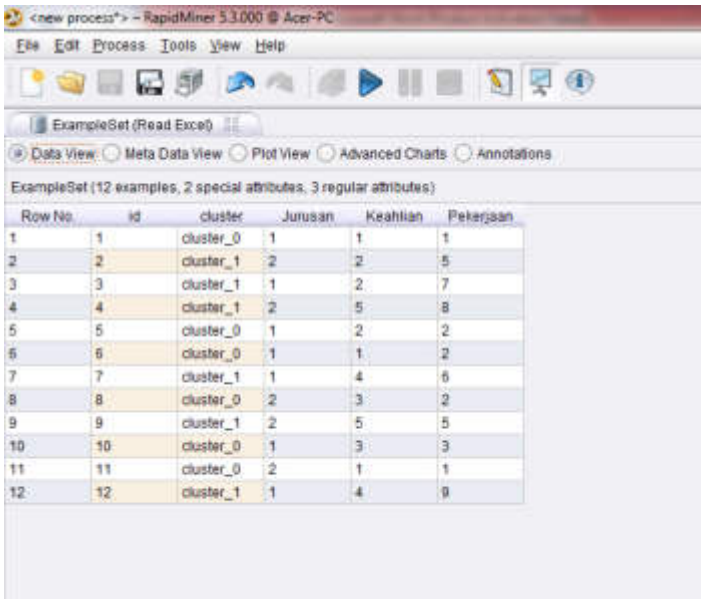


Figure 4. Data Input Process

B. **Results of data testing**

Testing was carried out using the rapid miner application, the test results are as shown in Figure 5 below:



Row No.	id	cluster	Jurusan	Keahlian	Pekerjaan
1	1	cluster_0	1	1	1
2	2	cluster_1	2	2	5
3	3	cluster_1	1	2	7
4	4	cluster_1	2	5	8
5	5	cluster_0	1	2	2
6	6	cluster_0	1	1	2
7	7	cluster_1	1	4	6
8	8	cluster_0	2	3	2
9	9	cluster_1	2	5	5
10	10	cluster_0	1	3	3
11	11	cluster_0	2	1	1
12	12	cluster_1	1	4	9

Figure 4. Data Input Process

From the test results, the results are the same as the manual calculation process where cluster 0 in the rapid miner application is the same as cluster 1 (C1) in manual calculation, namely data 1, data 5, data 6, data 8, data 10, and data 11 .

Cluster 1 is the same as cluster 2 (C2) with members data 2, data 3, data 4, data 7, data 9, and data 12.

4. CONCLUSION

The K-Means Clustering method succeeded in grouping alumni data according to the groups or clusters formed. The process of the k-means clustering algorithm obtains the results of grouping data to determine the linearity of alumni's work according to their field of knowledge with the results obtained that 50% of alumni are linear with fields of work and 50% of alumni are non-linear with occupations.

That isCluster 1 (C1) with cluster centers (1,333, 1,833 and 1,833) which are defined as Linear clusters namely alumni001, alumni005, alumni006, alumni008, alumni010, alumni011. Cluster 2 (C2) with cluster centers (1,500, 3,667 and 6,667) which is a non-linear cluster, where there are 6 data included in this cluster namely alumni002, alumni003, alumni004, alumni007, alumni009, alumni012.

REFERENCES

[1] T. Suprawoto, "Klasifikasi data mahasiswa menggunakan metode k- means untuk menunjang pemilihan strategi pemasaran," vol. 1, no. 1, pp. 12–18, 2016.
[2] R. Dalam and K. Baru, "Analisis Clustering K-Means Pada Pengelompokkan Hasil Tracer Study Sebagai Media Informasi Dalam Pengembangan Kurikulum Program Studi," vol. 3, 2019.
[3] K. Blitar, J. Timur, K. Kunci, K. K. Clustering, and K. S. Coeficient, "Penerapan Algoritma K-Means Clustering Untuk Menentukan Linieritas Pekerjaan Alumni Berdasarkan Tracer Study," no. September, pp. 3265–3281, 2022.
[4] U. U. Indonesia, S. Febrianti, L. Fitria, U. Samudra, L. Lama, and L. City, "PENERAPAN METODE K – MEANS CLUSTERING TERHADAP ALUMNI BERDASARKAN KUESIONER TRACER STUDY APPLICATION OF METHOD K – MEANS CLUSTERING TO ALUMNI BASED," vol. 7, no. 2, pp. 117–122, 2021.
[5] B. W. Nugraha, A. Mahmudi, F. S. Wahyuni, and F. T. Industri, "PENERAPAN METODE K-MEANS UNTUK PENGELOMPOKAN TINGKAT MALANG," vol. 5, no. 2, pp. 684–692, 2021.
[6] D. P. M and A. Fadlil, "Penerapan Clustering K-Means untuk Pengelompokan Tingkat Kepuasan Pengguna Lulusan Perguruan Tinggi," vol. 6, pp. 1693–1700, 2022, doi: 10.30865/mib.v6i3.4191.
[7] Y. C. Jimmy, "Perancangan Model Predıksı Performa Akademık Mahasıswa Menggunakan Algorıtma K - Means Clusterıng (Studı Kasus : Unıversıtas Xyz)," vol. 1, no. 1, pp. 643–649, 2021.
[8] R. Muktiadi and A. Y. Badharudin, "Metode K-Means untuk Mengelompokkan Alumni Berdasarkan Waktu Mencari Pekerjaan," vol. 16, no. 1, pp. 83–92, 2019.
[9] [9] W. Lestari, "Clustering Data Mahasiswa Menggunakan Algoritma K-Means Untuk Menunjang Strategi Promosi (Studi Kasus : STMIK Bina Bangsa Kendari)," vol. 4, no. 2, pp. 35–48, 2019.
[10] M. K. K-means, "Pokok Pembahasan".
[11] P. K. Clustering and S. Promosi, "LAPORAN AKHIR PENELITIAN Penerapan," no. November, 2022.
[12] I. Sumadikarta and E. Abeiza, "PENERAPAN ALGORITMA K-MEANS PADA DATA MINING UNTUK MEMILIH PRODUK DAN PELANGGAN POTENSIAL (Studi Kasus : PT Mega Arvia Utama)," *J. Satya Inform.*, no. 1, pp. 1–12, 2014.
[13] Z. Zulham and B. S. Hasugian, "Pengelompokan Siswa Dalam Menentukan Penerima Beasiswa Berdasarkan Prestasi Akademik Dengan Algoritma K-Means," *War. Dharmawangsa*, vol. 16, no. 3, pp. 231–241, 2022, doi: 10.46576/wdw.v16i3.2220.
[14] [S. N. Br Sembiring, H. Winata, and S. Kusnasari, "Pengelompokan Prestasi Siswa Menggunakan Algoritma K-Means," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 1, no. 1, p. 31, 2022, doi: 10.53513/jursi.v1i1.4784.
[15] F. Rini, N. Kahar, and Juliana, "Penerapan Algoritma K-Means Pada Pengelompokan Data Siswa Baru Berdasarkan Jurusan Di Smk Negeri 1 Kota Jambi Berbasis Web," *Semin. Nas. APTIKOM*, pp. 94–99, 2016.