International Journal of Computer Sciences and Mathematics Engineering Vol. 3 No. 1 (2024)
Published by: CV. HAWARI

International Journal of Computer Sciences and Mathematics Engineering
Journal homepage: www.ijecom.org

Analysis and Exploration of Clustering Algorithms for New Student Segmentation

Langgeng Restuono[1], Andysah Putera Utama Siahaan[2], Rian Farta Wijaya[3], Zulham Sitorus[4], Muhammad Iqbal[5]
[1,2,3,4,5]Department of Master of Information Technology, Universitas Pembangunan Panca Budi, Indonesia

Article Info

Article history:

Received May, 2024
Revised May, 2024
Accepted May, 2024

Keywords:

Clustering,
K-Means,
Centroid Initialization,
Random Centroids,
Manual Centroids,
Analysis and Interpretation

ABSTRACT

Clustering analysis is a crucial technique in data processing and pattern understanding. In this study, we compare the clustering results using the k-Means algorithm with two different approaches to centroid initialization: random centroids and manual centroids. The dataset consists of three observed variables. The analysis results indicate significant differences in centroid placement and cluster formation between the two approaches. The random centroid approach yields three clusters with centroids located at different coordinates: Cluster 1 [1.76, 2.5, 10.88], Cluster 2 [1.60, 1.87, 2.23], and Cluster 3 [1.64, 1.568, 15.88]. On the other hand, the manual centroid approach generates three clusters with centroids manually specified: Cluster 1 [1.64, 1.81, 14.84], Cluster 2 [1.61, 1.901, 2.04], and Cluster 3 [1.75, 1.7, 6.8]. The analysis and interpretation of these differences highlight the sensitivity of the k-Means algorithm to centroid initialization. The implications of these findings provide insights into the importance of selecting the appropriate initialization method in clustering analysis to ensure consistent and meaningful results. This research makes a significant contribution to understanding the factors influencing clustering results and can serve as a guide for researchers and practitioners in choosing clustering approaches that are suitable for their data and analytical goals.

Corresponding Author:

Langgeng Restuono1,
Department of Master of Information Technology,
Universitas Pembangunan Panca Budi,
Medan, Indonesia
Email: lsntl@ccu.edu.tw

1. INTRODUCTION

Institutions of higher education, especially in the field of Information Technology, are faced with increasingly complex challenges in understanding and meeting the needs of new students. New students often have diverse academic backgrounds, interests, and personal needs. Therefore, effective

ISSN 2962-4274

educational management requires an approach that can understand this diversity to enhance the academic experience and success of students.

Segmentation of new students is crucial in the context of higher education to understand their needs and preferences. With appropriate segmentation, universities can develop more effective enrollment strategies, improve student retention, and create a supportive academic environment. Therefore, the use of clustering algorithms in identifying patterns and groups of new students can be a very important initial step.

In the digital era, the use of technology such as data mining and data analysis has become one of the most effective approaches in higher education management. By leveraging this technology, educational institutions can optimize decision-making processes, improve operational efficiency, and provide better services to students.

The increasing diversity in academic backgrounds, cultures, and interests among new students adds complexity to managing higher education. This demands a more personalized and tailored approach to meet individual needs. By applying advanced segmentation analysis, institutions can identify underlying patterns in student preferences and adjust their strategies according to different needs.

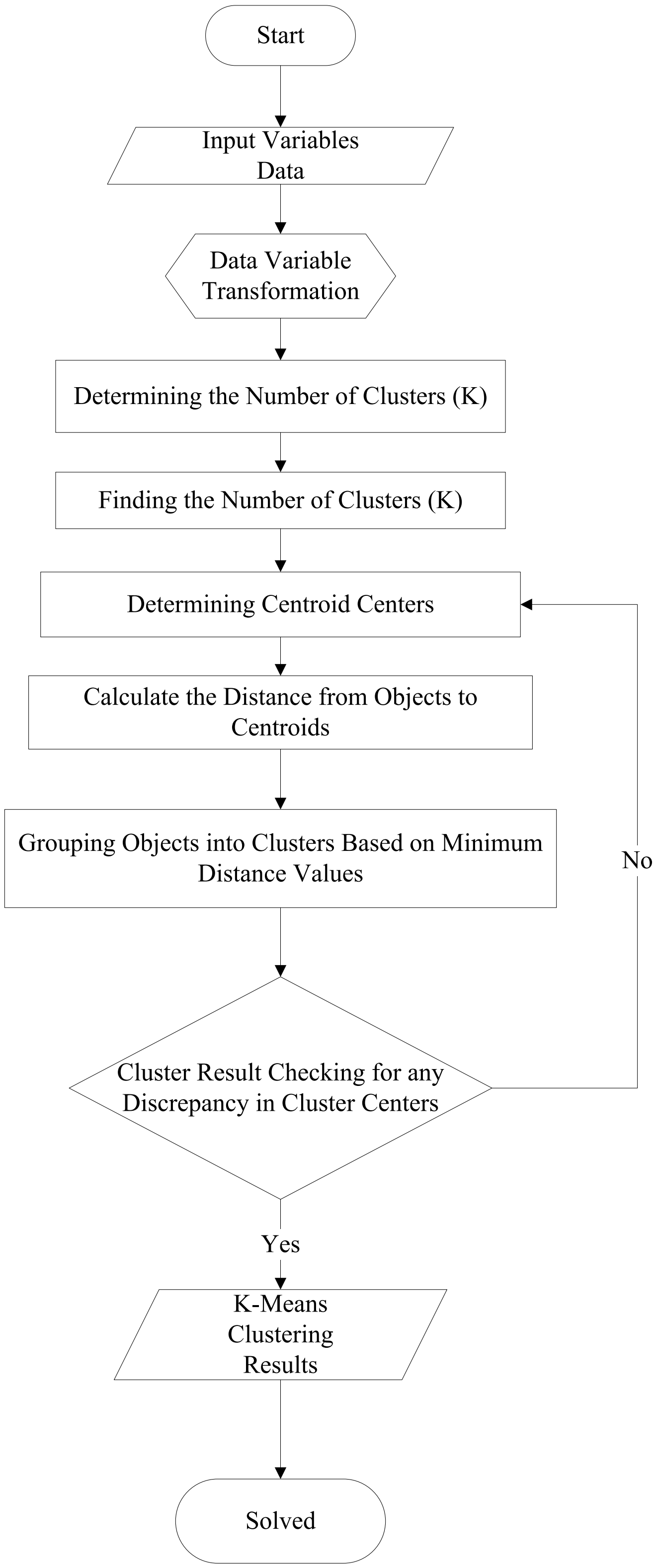
Intense competition among universities to attract new students adds urgency to the use of data analysis techniques. By understanding patterns of new student admissions and trends in the higher education market, universities can develop smarter and more efficient marketing strategies to attract the right prospective students.

In this context, this research aims to leverage the power of clustering algorithms to improve segmentation of new students. Better segmentation can assist institutions in devising more targeted strategies for new student integration, offering appropriate self-development programs, and aligning support services. However, this research limits its analysis to the data clustering stage only.

2. RESEARCH METHOD

Designing the Student Data Clustering Process, utilizing STMIK Kaputama student data. The required data for the clustering analysis process consists of pure database results, taking input variables such as school origin, parental occupation, and parental income (monthly income of the student's parents). These data serve as input, with the total count of students based on these factors serving as output. The stages of application development in student data clustering can be implemented into clustering calculations with the following flowchart:

ISSN 2962-4274



Description:
1. Prior to clustering, transform the data variables.

ISSN 2962-4274

2. Determine the number of clusters (K) and centroid centers as input, whether it's 2 clusters or more.
3. Set the centroid centers randomly.
4. Calculate the distance using Euclidean Distance.
5. Group objects based on the smallest distance calculation.
6. Repeat the process until no objects change clusters. If there are changes, repeat the steps of centroid determination, distance calculation, and grouping until there's no difference.
7. Output the results of K-Means clustering.
8. Finished.

Here is the process of K-Means clustering based on data from STMIK Kaputama:

Table 1: STMIK Kaputama Student Data for Clustering

No.	Name	School Origin	Parental Income/Month	Parental Occupation
1	Abdi Guna Setiawan	Esa Prakarsa	Rp. 2000.000 - Rp. 4.999.999	PNS
2	Ade Aprilia	SMA Swasta Persiaoan	Rp. 2.000.000-Rp. 4.999.999	PNS
3	Alifia Nazwa	MAN Binjai	Rp. 1.000.000 - Rp. 1.999.999	Tukang Tambal Ban
4	Alliya Dwi Ambarwati	SMK Yayasan Pendidikan Harapan Bangsa	Rp. 1.000.000-Rp. 1.999.999	Karyawan Swasta
5	Anastasya Viola Putri	SMA Swasta Persiapan Stabat	<Rp.500.000	Karyawan BUMN
6	Andika Riady Syahputra	SMK Putra Anda Binjai	Rp. 5.000.000 - Rp. 20.000.000	Karyawan Swasta
7	Anisa Herlina Putri Br Sembiring	Nurul Furqoon Binjai	Rp. 500.000 - Rp 999.999	Wiraswasta
8	Aqil Wahyu Pratama	SMA Negeri 1 Kuala	Rp. 1.000.000 - Rp. 1.999.999	Karyawan Swasta
9	Aulia Putri Padillah	SMA Swasta Persiapan Stabat	Rp. 1.000.000 - Rp. 1.999.999	Supir
10	Ayu Assyifa	SMAN 2 Binjai	Rp.2.000.000 - Rp.4.999.999	PNS
11	Cahaya Kamila	SMAN 4 Binjai	Rp. 1.000.000 - Rp. 1.999.999	Wiraswasta
12	Cinta Davita	Putra Anda Binjai	Rp. 1.000.000 - Rp. 1.999.999	Wiraswasta
13	Citra Pratiwi	SMA Negeri 1 Salapian	<Rp. 500.000	Petani
14	Deah Ajeng Agistira	SMA Negeri 5 Binjai	Rp. 2.000.000 - Rp. 4.999.999	Kepolisian RI (POLRI)
15	Deni Pratama	SMK Negeri 2 Binjai	<Rp. 500.000	Karyawan Swasta
16	Dianova Dwi Syafitri	MA Aisyiyah Binjai	Rp. 1.000.000-Rp. 1.999.999	Wiraswasta
17	Dito Oktama Putra	SMA Negeri 6 Binjai	Rp. 500.000 - Rp 999.999	Wirausaha
18	Dwi Irfan Hafiz	SMA Negeri 6 Binjai	Rp. 2.000.000 - Rp. 4.999.999	Polri
19	Dyo Alfattah	SMA Swasta Paba	Rp. 2.000.000 - Rp. 4.999.999	Polri
20	Edi Pindo Sitepu	SMA Negeri 1 Salapian	<Rp. 500.000	Petani
21	Ella Aisia	SMA N 3 Binjai	Rp. 1.000.000 - Rp. 1.999.999	Wiraswasta

ISSN 2962-4274

No.	Name	School Origin	Parental Income/Month	Parental Occupation
22	Elli Nurma Wati	Esa Prakarsa	Rp. 500.00-Rp. 999.999	Wiraswasta
23	Elsa Damayanti	SMK Swasta Al Wasliyah Stabat	<Rp. 500.000	Pensiunan Karyawan Swasta
24	Elvira Prananda	SMK N 1 Stabat	Rp. 1.000.000 - Rp. 1.999.999	Wiraswasta
25	Eron Garfil	SMAN 6 Binjai	Rp. 500.00-Rp. 999.999	Wiraswasta
26	Nazwa Intan Sari Br. Sitepu	Madrasah Aliyah Negeri Binjai	Rp. 1.000.000 - Rp. 1.999.999	Buruh
27	Vannisa Zahara	SMKN 1 Binjai	Rp. 1.000.000 - Rp. 1.999.999	Pegawai BUMN
28	Alya Velisia	SMA Esa Prakarsa	Rp. 1.000.000 - Rp. 1.999.999	Karyawan Swasta
29	Damaiyanti	SMKS Setia Budi Binjai	Rp. 0 – 999.999	DLL
30	Diana Cahaya Putri	SMA Eka Prakarsa	Rp. 500.000 - Rp 999.999	Wiraswasta

To process the data above using the K-Means Clustering method, the nominal and non-nominal data types such as school origin, parental occupation, and parental income need to be initialized into numerical form. The student data grouping can be expressed in independent variables, namely School Origin (X), Parental Income (Y), and Parental Occupation (Z).

Table 2: School Initialization

Numeric Code	School Origin
1	SMA/SMK Negeri
2	SMA/SMK Swasta
3	MAN
4	MAS

Table 3: Initialization of Socioeconomic Status Criteria (Parental Income)

Numeric Code	Income Range (IDR/Month)
1	Rp. 0 – 999.999
2	Rp. 1.000.000 – 1.999.999
3	Rp. 2.000.000 – 4.999.000
4	Rp. 5.000.001 – 7.000.000

Table 4: Initialization of Parental Occupation

Numeric Code	Occupation
1	Pegawai Negeri Sipil (PNS)
2	Wiraswasta
3	Petani
4	Nelayan
5	Pedagang
6	Kuli Bangunan
7	Supir
8	Security
9	Pensiunan PNS
10	Pesiunan BUMN
11	BUMN
12	TNI / POLRI
13	Pensiunan TNI / POLRI

ISSN 2962-4274

14	Buruh Haria Lepas
15	Karyawan Swasta
16	Wirausaha
17	Dan Lain-Lain

Table 5: Transformed Data Based on Encoding

No.	Name	School Origin	Parental Income/Month	Parental Occupation
1	Abdi Guna Setiawan	2	3	1
2	Ade Aprilia	2	3	1
3	Alifia Nazwa	3	2	17
4	Alliya Dwi Ambarwati	2	2	15
5	Anastasya Viola Putri	2	1	11
6	Andika Riady Syahputra	2	4	15
7	Anisa Herlina Putri Br Sembiring	4	1	2
8	Aqil Wahyu Pratama	1	2	15
9	Aulia Putri Padillah	2	2	7
10	Ayu Assyifa	1	3	1
11	Cahaya Kamila	1	2	2
12	Cinta Davita	2	2	2
13	Citra Pratiwi	1	1	3
14	Deah Ajeng Agistira	1	3	12
15	Deni Pratama	1	1	15
16	Dianova Dwi Syafitri	4	2	2
17	Dito Oktama Putra	1	1	16
18	Dwi Irfan Hafiz	1	3	12
19	Dyo Alfattah	2	3	12
20	Edi Pindo Sitepu	1	1	3
21	Ella Aisia	1	2	2
22	Elli Nurma Wati	2	1	2
23	Elsa Damayanti	2	1	15
24	Elvira Prananda	1	2	2
25	Eron Garfil	1	1	2
26	Nazwa Intan Sari Br. Sitepu	3	2	14
27	Vannisa Zahara	1	2	11
28	Alya Velisia	2	2	15
29	Damaiyanti	2	1	17
30	Diana Cahaya Putri	2	1	2

These transformed values can be used for further analysis using the K-Means clustering algorithm. Let me know if you need any more assistance!

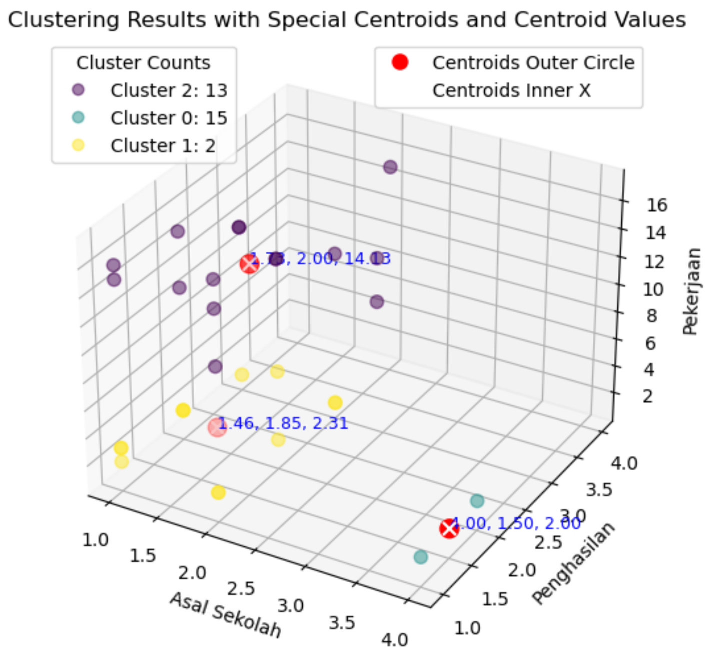


Figure 1. Cluster graph based on the calculations performed.

Explanation of the Graph:
From 30 data, 3 groups were obtained. Cluster 1 consists of 15 student data, cluster 2 consists of 2 student data, and cluster 3 consists of 13 student data, with the largest group obtained in cluster 1.

1. Cluster 1: Consists of 15 Student Data
   - School Origin: 1.73 (SMA/SMK Private)
   - Parental Income: 2 (Income Range: Rp. 1,000,000 - Rp. 1,999,999)
   - Parental Occupation: 14.13 (Private Employee)

It can be observed that in cluster 1, many students from STMIK Kaputama are from private high schools (SMA/SMK Swasta) with socioeconomic status (Parental Income) ranging from Rp. 1,000,000 to Rp. 1,999,999 and Parental Occupation as Private Employees.

2. Cluster 2: Consists of 2 Student Data
   - School Origin: 4 (SMA)
   - Parental Income: 1.5 (Income Range: Rp. 1,000,000 - Rp. 1,999,999)
   - Parental Occupation: 2 (Entrepreneur)

It can be observed that in cluster 2, many students from STMIK Kaputama are from public high schools (SMA) with socioeconomic status (Parental Income) ranging from Rp. 1,000,000 to Rp. 1,999,999 and Parental Occupation as Entrepreneurs.

3. Cluster 3: Consists of 13 Student Data
   - School Origin: 1.46 (SMA/SMK Public)
   - Parental Income: 1.85 (Income Range: Rp. 0 - Rp. 1,999,999)
   - Parental Occupation: 2.31 (Entrepreneur)

It can be observed that in cluster 3, many students from STMIK Kaputama are from public high schools (SMA/SMK Negeri) with socioeconomic status (Parental Income) ranging from Rp. 0 to Rp. 1,999,999 and Parental Occupation as Entrepreneurs.

3. RESULTS AND DISCUSSIONS

The steps taken for calculating student data using the clustering method with the K-means algorithm aim to generate new knowledge about the number of groups based on school origin, parental income, and parental occupation data of STMIK Kaputama students. This allows us to determine the closest relationship between student data groups.

After the data is imported into Python and processed using the specified syntax, the clustering results are divided into three groups based on the closest distance from the centroid. This data includes the variables School Origin (X), Income (Y), and Parental Occupation (Z), which can be seen in the following table:

Table 6: Group Determination Results

No	X	Y	Z	Group
1	2	3	1	2
2	2	3	1	2
3	3	2	17	3
4	2	2	15	3
5	2	1	11	1
6	2	4	15	3
7	4	1	2	2
8	1	2	15	3
9	2	2	7	1
10	1	3	1	2
11	1	2	2	2
12	2	2	2	2
13	1	1	3	2
14	1	3	12	1
15	1	1	15	3

No	X	Y	Z	Group
16	4	2	2	2
17	1	1	16	3
18	1	3	12	1
19	2	3	12	1
20	1	1	3	2
21	1	2	2	2
22	2	1	2	2
23	2	1	15	3
24	1	2	2	2
25	1	1	2	2
26	3	2	14	3
27	1	2	11	1
28	2	2	15	3
...	...	...	...	...
357	1	1	17	3

3.1. Static Centroids

Static centroids refer to fixed initial positions assigned to the centroids at the beginning of the clustering algorithm and remain unchanged throughout the iteration process. In other words, the centroid positions are set statically and do not adapt or update based on the data during the clustering process

ISSN 2962-4274

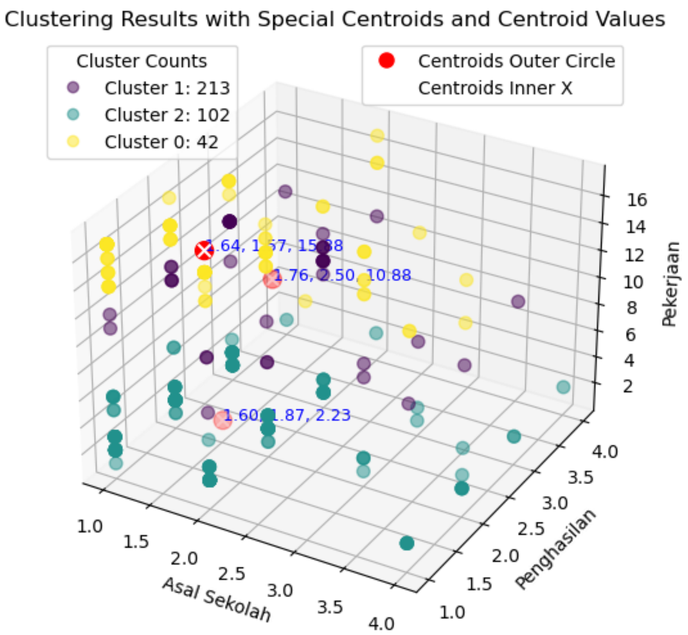


Figure 2. Cluster graph based on the calculations performed Static Centroids

Explanation:

From 357 student data, 3 clusters are obtained. Cluster 1 consists of 42 data, cluster 2 consists of 213 data, and cluster 3 consists of 102 data. Below are the descriptions of the cluster centroids on the graph:

1. Group 1 / Cluster 1: 1 (1.76) 3 (2.50) 12 (10.88)
   It can be observed that in cluster 1, the group of student data originates from public high schools (SMA/SMK Negeri) with parental income ranging from Rp. 2,000,000 to Rp. 4,999,999 and with parents employed as military/police officers.
2. Group 2 / Cluster 2: 1 (1.60) 2 (1.87) 2 (2.23)
   It can be observed that in cluster 2, the group of student data originates from public high schools (SMA/SMK Negeri) with parental income ranging from Rp. 1,000,000 to Rp. 1,999,999 and with parents employed as entrepreneurs.
3. Group 3 / Cluster 3: 1 (1.64) 1 (1.57) 17 (15.88)
   It can be observed that in cluster 3, the group of student data originates from public high schools (SMA/SMK Negeri) with parental income ranging from Rp. 0 to Rp. 999,999 and with parents employed in other occupations.

These descriptions provide insights into the characteristics of each cluster based on school origin, parental income, and parental occupation.

3.2. Dynamic Centroids

Dynamic centroids, on the other hand, involve centroid positions that can change or adapt during the clustering iteration process. These centroids are initialized randomly or based on certain criteria but are allowed to move towards the center of their respective clusters as the algorithm progresses. Dynamic centroids enable the algorithm to better capture the underlying patterns in the data and adjust to the distribution of the data points. This flexibility can lead to more accurate clustering results, especially in situations where the data distribution is complex or changes over time.

ISSN 2962-4274

After importing the data into Python using the specified syntax, based on the distance from the manually determined cluster centroids, the selected centroids are as follows:
Centroid 1 = (1, 2, 17)
Centroid 2 = (2, 3, 12)
Centroid 3 = (4, 4, 12)
The results of the group calculation are divided into 3 groups with school origin (X), income (Y), and occupation (Z) data.

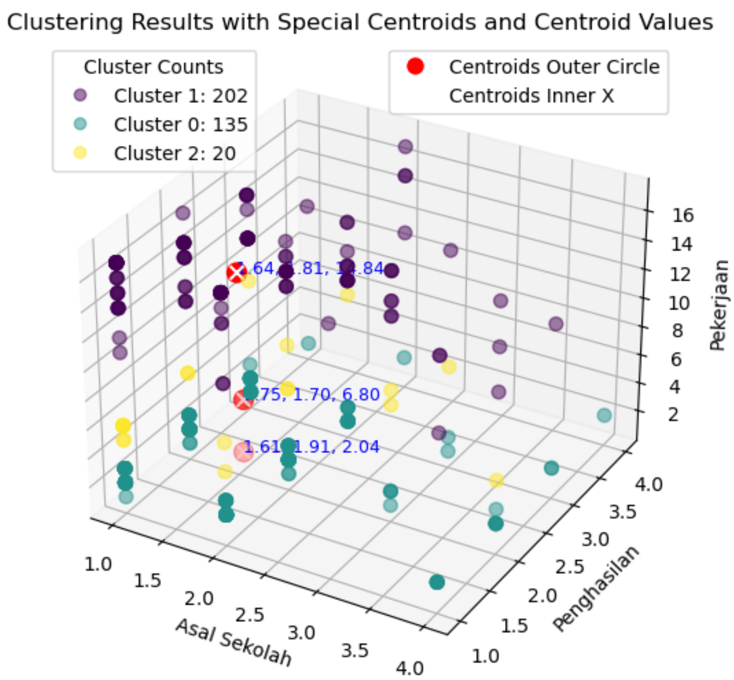


Figure 3. Cluster graph based on the calculations performed Dynamic Centroids

Explanation:

From 357 student data, 3 clusters were obtained using manually determined centroids, where cluster 1 consists of 135 data, cluster 2 consists of 202 data, and cluster 3 consists of 20 data. Below are the descriptions of the cluster centroids on the graph along with the groups:

1. Group 1 / Cluster 1 with values 1 (1.64) 1 (1.81) 17 (14.84)

   It can be observed that in group 1, there are student data with an average origin from public high schools (SMA/SMK Negeri) with an average parental income ranging from Rp. 0 to Rp. 999,000 and an average occupation categorized as others.

2. Group 2 / Cluster 2 with values 1 (1.61) 2 (1.91) 2 (2.5)

   It can be observed that in group 2, there are student data with an average origin from public high schools (SMA/SMK Negeri) with an average parental income ranging from Rp. 1,000,000 to Rp. 1,999,999 and an average occupation categorized as entrepreneurs.

3. Group 3 / Cluster 3 with values 1 (1.75) 1 (1.7) 6 (6.8)

   It can be observed that in group 3, there are student data with an average origin from public high schools (SMA/SMK Negeri) with an average parental income ranging from Rp. 0 to Rp. 999,999 and an average occupation categorized as construction workers.

ISSN 2962-4274

From the initial to final analysis with cluster 3 centroids with random and manually determined cluster values, there are differences observed.

4. CONCLUSION

The results of clustering analysis using the k-Means algorithm with a static centroid approach produced three main clusters with the following centroids:

- Cluster 1: [1.76, 2.5, 10.88]
- Cluster 2: [1.60, 1.87, 2.23]
- Cluster 3: [1.64, 1.56, 15.88]

Meanwhile, in the clustering approach with dynamic centroids, the following results were obtained::

- Cluster 1: [1.64, 1.81, 14.84]
- Cluster 2: [1.61, 1.91, 2.04]
- Cluster 3: [1.75, 1.7, 6.8]

A comparison between the clustering results using random centroids and manual centroids shows significant differences in centroid placement and cluster formation. This indicates that centroid initialization affects the final clustering results. Although both approaches resulted in three clusters, the centroid locations and data distributions within each cluster can vary substantially.

It is important to note that the choice of centroid initialization method can influence clustering results. This study provides insights into the sensitivity of the k-Means algorithm to centroid initialization. These results underscore the importance of selecting the appropriate initialization method in clustering analysis to ensure consistent and meaningful results. This can be a critical consideration for researchers and practitioners in choosing the clustering approach that best fits their data and analysis goals.

REFERENCES

[1] Budiman, Ramdani. "Penerapan Data Mining Untuk Menentukan Lokasi Promosi Penerimaan Mahasiswa Baru Pada Universitas Banten Jaya (Metode K-Means Clustering)." *ProTekInfo (Pengembangan Riset dan Observasi Teknik Informatika)* 6 (2019): 6-14.

[2] Bellanov, Agrienta. "K-Means Clustering Analysis Untuk Menentukan Strategi Promosi Kampus." *Jurnal Teknik Industri: Jurnal Hasil Penelitian dan Karya Ilmiah dalam Bidang Teknik Industri* 9.1 (2023): 259-268.

[3] Khusnuliawati, Hardika, and Dhian Riskiana Putri. "Identifikasi Segmen Pasar Mahasiswa Perguruan Tinggi Menggunakan Analisis Klaster Berdasarkan Variabel Psikografis." *Risenologi* 6.1b (2021): 44-49.

[4] Annizar, Anas Ma'ruf, and Miftah Arifin. "Perbedaan Prestasi Belajar Mahasiswa Ditinjau dari Jalur Seleksi Masuk Perguruan Tinggi." *SAP (Susunan Artikel Pendidikan)* 5.3 (2021).

[5] Muhima, Rani Rotul, et al. *Kupas Tuntas Algoritma Clustering: Konsep, Perhitungan Manual, dan Program*. Penerbit Andi, 2022.

[6] Arsyad, Aisyah Tiar, and Hanny Nurlatifah. "Penerapan k-means clustering dalam menentukan Strategi promosi Universitas Al Azhar Indonesia." (2022).

[7] Burk, Scott, and Gary D. Miner. *It's All Analytics!: The Foundations of Al, Big Data and Data Science Landscape for Professionals in Healthcare, Business, and Government*. CRC Press, 2020.

[8] Sun, Zhaohao. "Data, Analytics, and Intelligence." *Journal of Computer Science Research* 5.4 (2023): 43-57.

[9] Ajimotokan, Habeeb Adewale. *Research Techniques: Qualitative, Quantitative and Mixed Methods Approaches for Engineers*. Springer Nature, 2022.

[10] Mueller, Jennifer J., et al. *Understanding research in early childhood education: Quantitative and qualitative methods*. Taylor & Francis, 2024.

ISSN 2962-4274

[11] Amin, Nur Fadilah, Sabaruddin Garancang, and Kamaluddin Abunawas. "Konsep Umum Populasi dan Sampel dalam Penelitian." *PILAR* 14.1 (2023): 15-31.

[12] Abriyanto, Arif, and Natalia Damastuti. "SEGMENTASI MAHASISWA DENGAN 'UNSUPERVISED'ALGORITMA GUNA MEMBANGUN STRATEGI MARKETING PENERIMAAN MAHASISWA." *Insand Comtech: Information Science and Computer Technology Journal* 4.2 (2019).

[13] Bahri, S. (2018). Metodologi Penelitian Bisnis Lengkap dengan teknik Pengolahan Data SPSS. Yogyakarta: CV ANDI OFFSET.

[14] Suhanda, Yogasetya, Ike Kurniati, and Siti Norma. "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik." *Jurnal Teknologi Informatika dan Komputer* 6.2 (2020): 12-20.

[15] Sujarweni, Wiratna. (2020). Metodologi Penelitian Bisnis & Ekonomi. Yogjakarta

[16] Abriyanto, Arif, and Natalia Damastuti. "SEGMENTASI MAHASISWA DENGAN 'UNSUPERVISED'ALGORITMA GUNA MEMBANGUN STRATEGI MARKETING PENERIMAAN MAHASISWA." *Insand Comtech: Information Science and Computer Technology Journal* 4.2 (2019).

[17] Hendryadi, Tricahyadinata, I., & Zannati, R. (2019). Metode Penelitian: Pedoman Penelitian Bisnis dan Akademik. Jakarta: Lembaga Pengembagan Manajemen dan Publikasi Imperium (LPMP Imperium)

[18] Wu, Junjie. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media, 2012.

[19] Rijali, Ahmad. "Analisis data kualitatif." *Alhadharah: Jurnal Ilmu Dakwah* 17.33 (2019): 81-95.