Analisis K-Means Clustering pada Data Sepeda Motor

Rozzi Kesuma Dinata[1], Safwandi[2], Novia Hasdyna[3], Nur Azizah[4]
[1,2,4] Program Studi Teknik Informatika, Universitas Malikussaleh
[3]Program Studi Teknik Informatika, Universitas Islam Kebangsaan Indonesia
[1]rozzi@unimal.ac.id , [2]safwandi@unimal.ac.id , [3]noviahasdyna@gmail.com, [4]nurazizah@gmail.com

ABSTRACT

K-Means is a data mining algorithm that can be used to grouping or clustering data. This research using k-means for clustering the data of motorcycle based on consumer needs. The dataset used in this research is Honda and Yamaha motorcycle which taken from the dialers in Dewantara District, Aceh. The data tested by grouping 300 data of motorcycle with different attributes into 3 clusters, which are cheap, normal, and expensive. The distribution of the data we separate it using 45 data in 15 times of test. Each test used 3 different data randomly selected on each test. To calculate the distance of each motorcycle data that have been inputted to each centroid, we used the Euclidean Distance formula. Data in this cluster system can be used as a recommendation for users in selecting the motorcycle that they interest the most. The results of the performance on each test finished in 15 times shown that the average value of Precision by 76%, Recall by 76% and the accuracy by 81%.

Keyword: *Data Mining, K-Means, Clustering, Euclidean Distance*

1. Introduction

Proses ekstraksi data menjadi informasi yang sebelumnya belum tersampaikan, dengan proses data mining dan teknik yang tepat akan memberikan hasil yang optimal [2]. Data Mining merupakan serangkaian proses dalam pencarian pola, hubungan, penggalian nilai tambah dari data dan informasi yang berukuran besar berupa pengetahuan dengan tujuan menemukan hubungan dan menyederhanakan data agar diperoleh informasi yang mudah dipahami dan bermanfaat [3]. Salah satu teknik pengelompokan sepeda motor berdasarkan kebutuhan konsumen yang dapat digunakan dalam data mining adalah metode pengelompokan Clustering. Clustering adalah metode yang digunakan dalam data mining dengan cara kerjanya mencari data dan menglompokkan data yang mempunyai kemiripan karakteristik antara data satu dengan data lainnya yang telah diperoleh [4].

Menurut kategori kekompakan, pengelompokan terbagi menjadi dua, yaitu komplet dan parsial. Jika semua data dapat bergabung menjadi satu, maka dikatakan semua data kompak menjadi satu kelompok [5]. Metode Clustering yang dapat digunakan salah satunya adalah metode k-means karena K-Means merupakan salah satu algoritma dalam data mining yang bisa digunakan untuk melakukan pengelompokan atau Clustering suatu data yaitu data sepeda motor [6]. Tujuan pengelompokkan data ini adalah untuk meminimalkan fungsi objektif yang diatur dalam proses pengelompokan, pada umumnya berusaha meminimalkan variasi suatu kelompok dan memaksimalkan variasi antar kelompok [7].

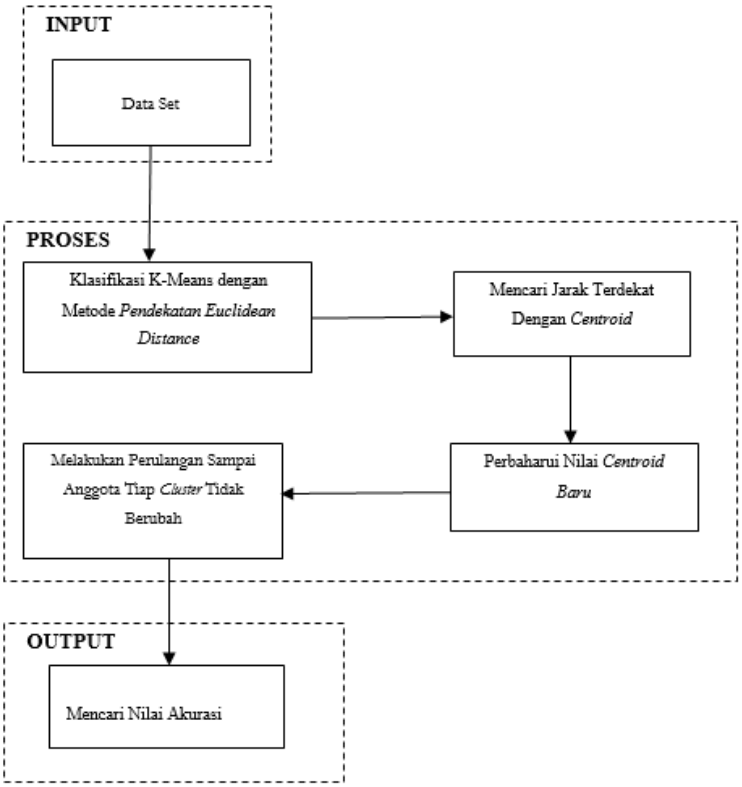
Penelitian tentang clustering data telah banyak dikembangkan oleh para peneliti. Ada beragam metode data mining yang telah diteliti, diantaranya seperti dalam peneliti yang telah diteliti oleh Putra et al. (2017) menggunakan *hadoop single node cluster* dengan metode K-Nearest Neighbor dalam klasifikasi sepeda motor berdasarkan karakteristik konsumen. Uji coba yang dilakukan menghasilkan merk sepeda motor dari data uji yang ditentukan [1]. Penelitian yang dilakukan oleh Purwadi (2018) menggunakan metode algoritma c4 untuk memprediksi pola pembelian sepeda motor pada showroom cv. viva mas motors [8].

Pada penelitian ini menerapkan k-means dalam pengelompokan data sepeda motor. Hasil dari penelitian ini berupa clustering yang memisahkan data menjadi 3 cluster, yaitu murah, standard, dan mahal. Data dalam cluster ini dapat menjadi rekomendasi bagi pengguna dalam menentukan pemilihan sepeda motor yang diinginkan. Adapun hasil analisis juga dapat berupa pengukuran performansi seperti *accuracy*, *recall* dan *precision*.

2. Research Method

2.1. Diagram Alir Penelitian

Penelitian ini terdiri dari 6 proses, seperti dalam diagram alir penelitian berikut.



Gambar 1. Digram Alir Penelitian

Penjabaran dari diagram alir k-means adalah:

1. Langkah pertama adalah memilih dataset yang akan digunakan dalam proses penelitian, dalam penelitian ini menggunakan dataset Sepeda Motor merek Honda dan Yamaha
2. Langkah selanjutnya pembagian dataset data training dan data testing. Setelah pembagian data menjadi 2 bagian selanjutnya melakukan pengujian dengan menggunakan Algoritma K-Means.
3. Langkah selanjutnya adalah mencari jarak terdekat dengan Centroid menggunakan rumus jarak Euclidian Distance.
4. Setelah itu memperbaharui nilai Centroid baru yang diperoleh dari rata-rata Cluster. Langkah terakhir yaitu mecari nilai akurasi dari setiap pengujian yang dilakukan.

2.2. Algoritma K-Means

K-Means adalah merupakan salah satu metode dalam data mining yang dapat mengelompokkan data atau *Clustering* sebuah data kedalam bentuk satu *cluster* atau lebih *cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data dengan karakteristik yang berbeda dikelompokkan ke dalam kelompok berbeda yang lainnya [9]. Sarwono mengemukakan secara detail, algoritma *K-Means* adalah sebagai berikut yaitu [10] :

1. Menentukan nilai k sebagai jumlah *cluster* yang ingin di bentuk.
2. Menentukan nilai acak atau random untuk pusat *cluster* awal centroid sebanyak k, untuk menghitung jarak setiap data input terhadap masing–masing *centroid* dengan menggunakan rumus jarak *Euclidean Distance* yaitu :

$$d(xi, \mu j) = \sqrt{\sum (xi - \mu j)^2} \qquad (1)$$

Dimana: xi = data kriteria
μj = *centroid* pada *cluster* ke-js

3. Mengelompokkan setiap data berdasarkan kedekatannya dengan *centroid* atau mencari jarak terkecil.
4. Memperbaharui nilai *centroid* baru, nilai *centroid* baru di peroleh dari rata-rata *cluster* yang bersangkutan dengan menggunakan rumus yaitu :

$$\mu j(t+1) \quad = \frac{1}{Nsj} \sum_{j \in sj} xj \qquad (2)$$

Keterangan : μj(t+1) = *centroid* baru pada iterasi (t+1)
N*sj* = Data pada *cluster* Sj

5. Apabila data setiap *cluster* belum berhenti, lakukan perulangan dari langkah 2 hingga 5, sampai anggota tiap *cluster* tidak ada yang berubah.

2.3. Confusion Matrix

Confusion matrix melakukan pengujian untuk memperdiksikan suatu objek yang benar dan salah [11].

Tabel 1. Cofusion Matrix

Nilai prediksi	Nilai Aktual	
	TP	TN
	FP	TN

TP = *True* positif yang diklasifikasikan positif.
TN= *True* negatif yang diklasifikasikan negatif.
FP = *False* positif yang diklasifikasikan positif.
FN= *False* negatif yang diklasifikasikan negatif.

Rumus untuk perhitungan *confusion matrix* untuk menghitung *presicion, recall,* dan nilai *accuracy* dapat dijelaskan di bawah ini :

a. *Presicion* berguna untuk mengukur tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

$$Precision = \frac{TP}{(TP + FP)} \qquad (3)$$

b. *Recall* berguna untuk mengukur tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi, pada persamaan berikut

$$Recall = \frac{TP}{(TP + FN)} \qquad (4)$$

c. *Acuracy* berguna untuk mengukur suatu kinerja sebuah metode [12].

$$Acuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \qquad (5)$$

3. Result and Analysis

3.1. Dataset Sepeda Motor

Adapun penggunaan dataset sepeda motor dapat dilihat pada tabel berikut.

Tabel 1. Dataset Sepeda Motor

No	Nama	X1	X2	X3	X4	X5
1	Revo Fit	1	4	1	3	1
2	Revo Fit	1	2	1	3	1
3	Revo Fit	1	7	1	3	1
4	Revo X	1	8	1	3	3
5	Revo X	1	5	1	3	3
6	Blade 125 FI R	1	4	2	3	4
7	Blade 125 FI R	1	10	2	3	4
8	Blade 125 FI R	1	9	2	3	4
9	Supra X 125 SW	1	6	2	3	4
10	Supra X 125 SW	1	4	2	3	4

11	Supra X 125 SW	1	6	2	3	5
12	Supra X 125 SW	1	4	2	3	5
13	Supra X CW	1	4	2	9	5
14	Supra X CW	1	2	2	9	5
15	Supra X CW	1	11	2	9	5
…	…	…	…	…	…	…
…	…	…	…	…	…	…
300	Jupiter MX King	2	2	3	5	11

3.2. Data Testing

Data yang digunakan untuk proses pengujian dapat dilihat pada tabel berikut:

Tabel 2. Data *Testing*

Uji ke-	Data Uji
1	2, 87, 100
2	1, 6, 9
3	3, 14, 23
4	61, 81, 101
5	57, 66, 84
6	37, 47, 95
7	113, 139, 150
8	117, 137, 157
9	182, 190, 210
10	202, 220,240
11	120, 142, 161
12	124, 146, 153
13	257, 263, 276
14	183, 200, 213
15	268, 274, 280

3.3 Hasil Perhitungan Algoritma K-Means

Adapun hasil perhitungan jarak data ke-1 pada pengujian pertama dengan data uji masing-masing *cluster* yang ada di tabel 3 adalah:

$d_{(2.1)} = \sqrt{(1-1)^2+(2-4)^2+(1-1)^2+(3-3)^2+(1-1)^2}$
$= \sqrt{4}$
$= 2$

$d_{(87.1)} = \sqrt{(1-1)^2+(6-4)^2+(3-1)^2+(15-3)^2+(19-1)^2}$
$= \sqrt{476}$
$= 21{,}81742423$

$d_{(100.1)} = \sqrt{(1-1)^2+(5-4)^2+(3-1)^2+(18-3)^2+(38-1)^2}$
$= \sqrt{1599}$
$= 39{,}98749805$

Persamaan dengan perhitungan yang sama diterapkan pada 300 data untuk mendapatkan jarak tiap data pada masing-masing *cluster* seperti pada tabel 3.

Tabel 3. Jarak Pada Tiap *Cluster*

No	A2	A87	A100	Cluster
1	2	21,81742	39,9875	1
2	0	22,09072	40,0874	1
3	5	21,74856	40,02499	1
4	6,324555	20,19901	38,24918	1
5	3,605551	20,12461	38,13135	1
6	3,741657	19,33908	37,18871	1
7	8,602325	19,64688	37,51	1
8	7,681146	19,46792	37,38984	1
9	5,09902	19,23538	37,18871	1
10	3,741657	19,33908	37,18871	1

11	5,744563	18,46619	36,27671	1
12	4,582576	18,57418	36,27671	1
13	7,549834	15,3948	34,23449	1
14	7,28011	15,77973	34,35113	1
15	11,57584	16,06238	34,74191	1
…	…	…	…	…
…	…	…	…	…
300	10,44031	13,45362	30,13304	1

Setelah masing-masing data dihitung jaraknya untuk tiap *cluster*, langkah selanjutnya yaitu mengelompokkan data sesuai clusternya, kelompok *cluster* suatu data dihitung dari jarak terpendek dari data terhadap suatu *cluster.* Hasil pengelompokan dapat dilihat pada tabel 4.

Tabel 4. *Cluster* Dengan Jarak Terdekat

No	C1	C2	C3
1	*		
2	*		
3	*		
4	*		
5	*		
15	*		
16	*		
17		*	
19		*	
20		*	
21		*	
22		*	
23			*
144			*
145			*
…	…	…	…
…	…	…	…
300	*		

Penempatan data pada *cluster* dengan jarak terdekat ke *cluster* yaitu memperoleh data yang dekat dengan C_1=150, C_2=94, C_3=55. Setelah data dikelompokkan sesuai dengan clusternya, langkah selanjutnya adalah menghitung nilai *centroid* baru masing-masing *cluster* dengan menggunakan persamaan 2.

Perhitungan yang dilakukan untuk menghitung nilai *centroid* baru pada masing-masing *centroid* yaitu dengan menjumlahkan semua nilai pada setiap *cluster* yang sama dan membagikannya dengan jumlah data yang ada pada *cluster* tersebut. Hasil *centroid* baru dapat dilihat pada tabel 5.

Tabel 5. Nilai *Centroid* Baru

C1	**1,41059**	**6,03311**	**1,95364**	**4,45695**	**4,69536**
C2	1,67021	9,43617	3,07446	15,7766	15,4042
C3	1,23636	11,2	4,67272	22,1090	34,4727

Setelah nilai *centroid* baru dihitung, langkah selanjutnya adalah bandingkan dengan nilai *centroid* sebelumnya, jika nilainya sama maka proses iterasi dihentikan. Namun jika nilainya tidak sama langkah-langkah proses pengelompokan data diulangi kembali. Pada pengujian pertama ini datanya belum sama, datanya sama dan berhenti pada iterasi ke-4. Pada masing-masing pengujian data yang sudah sama tidak semua berhenti pada iterasi yang sama tetapi berhenti dibeda iterasi, untuk pemberhentian data pada iterasi berapa pada setiap pengujian bisa dilihat pada tabel 6 dibawah ini.

Tabel 6. Iterasi Pengujian

Uji ke-	Data Uji	Berhenti
1	2, 87, 100	di iterasi ke-4
2	1, 6, 9	di iterasi ke-4
3	3, 14, 23	di iterasi ke-6

4	61, 81, 101	di iterasi ke-6
5	57, 66, 84	di iterasi ke-9
6	37, 47, 95	di iterasi ke-7
7	113, 139, 150	di iterasi ke-6
8	117, 137, 157	di iterasi ke-6
9	182, 190, 210	di iterasi ke-7
10	202, 220,240	di iterasi ke-8
11	120, 142, 161	di iterasi ke-8
12	124, 146, 153	di iterasi ke-7
13	257, 263, 276	di iterasi ke-6
14	183, 200, 213	di iterasi ke-8
15	268, 274, 280	di iterasi ke-6

Hasil pengelompokan data ke 300 data sepeda motor dengan 15 kali pengujin dengan setiap pengujian menggunakan 3 *cluster* data disetiap masing-masing pengujian. Setelah dikelompokkan semua datanya, langkah berikutnya adalah menghitung nilai dengan *Confusion Matrix*.

Tabel 7. *Predicted Class*

Actual Class				
	Murah	Standar	Mahal	
Murah	107	0	0	107
Standar	50	82	1	133
Mahal	0	4	56	60
Rata-rata	157	86	57	300

Untuk menghitung nilai *Confusion Matrix* menggunakan nilai prediksi TP, FP,FN,TN.

Tabel 8. *Confusion Matrix*

	Class	**TP**	**FP**	**FN**	**TN**	**Total**
Confusion Matrik	Murah	107	50	0	143	300
	Standar	82	4	51	163	300
	Mahal	56	1	4	239	300

Langkah selanjutnya menghitung nilai *presicision*, *recall*, dan *accuracy,* masing-masing *class*, kemudian hitung nilai rata-rata semua *class*.

Tabel 9. Hasil *Confusion Matrix*

	Class	**Precision**	**Recall**	**Akurasi**
Confusion Matrik	Murah	0,681529	1	0,833333
	Standar	0,953488	0,616541	0,816667
	Mahal	0,982456	0,933333	0,983333
	Jumlah	87%	85%	88%

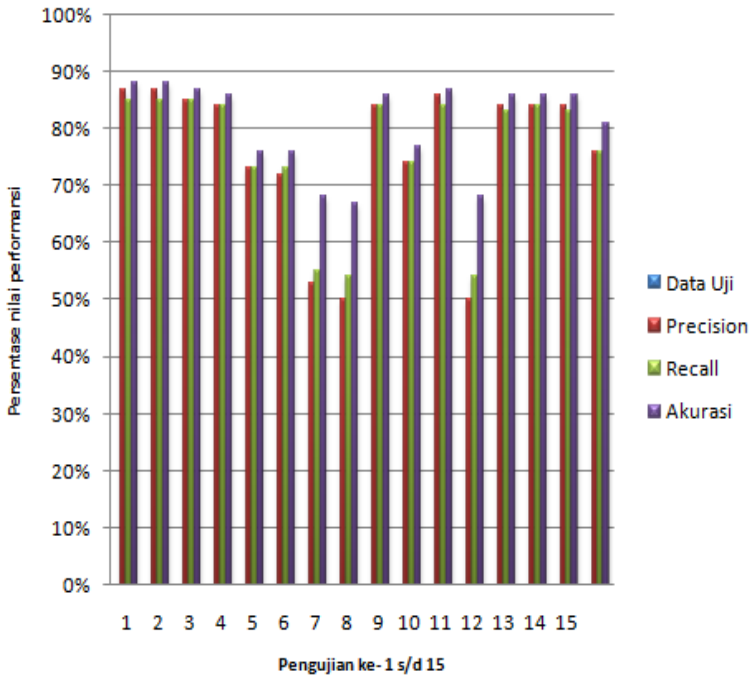
Setelah Hasil *Confusion Matrix* sudah selesai dicari semua, langkah selanjutnya yaitu menggabungkan semua nilai *Confusion Matrix* untuk mencari nilai rata-rata dari nilai *presicision*, *recall*, dan *accuracy,* dari semua ke-15 kali pengujian yang sudah dicari sebelumnya. Hasilnya dapat dilihat pada tabel 10 dibawah ini.

Tabel 10. Hasil *Confusion Matrix* untuk semua pengujian

Uji ke	**Data Uji**	**Precision**	**Recall**	**Akurasi**
1	2, 87, 100	87%	85%	88%
2	1, 6, 9	87%	85%	88%
3	3, 14, 23	85%	85%	87%
4	61, 81, 101	84%	84%	86%
5	57, 66, 84	73%	73%	76%
6	37, 47, 95	72%	73%	76%
7	113, 139, 150	53%	55%	68%
8	117, 137, 157	50%	54%	67%

9	182, 190, 210	84%	84%	86%
10	202, 220,240	74%	74%	77%
11	120, 142, 161	86%	84%	87%
12	124, 146, 153	50%	54%	68%
13	257, 263, 276	84%	83%	86%
14	183, 200, 213	84%	84%	86%
15	268, 274, 280	84%	83%	86%
Rata-Rata Performansi		76%	76%	81%

Bersasarkan tabel diatas, dapat dilihat bahwa pada clustering data sepeda motor dengan k-means dilakukan sebanyak 15 kali pengujian. Masing-masing pengujian dilakukan terhadap 3 data secara random. Hasil pengujian dalam bentuk pengukuran performansi k-means, berupa precision, recall dan akurasi. Seperti pada pengujian ke-1, data yang diambil adalah data ke-2,87,dan 100, dengan Nilai *precision* sebesar 87%, *recall* 85%, akurasi 88%. Pada pengujian ke-15, data yang diambil adalah data ke-268,274,dan 280, dengan Nilai *precision* sebesar 84%, *recall* 83%, akurasi 81%. Adapun nilai performansi dari metode *K-Means* dalam bentuk grafik bisa dilihat pada gambar dibawah ini.



Gambar 2. Grafik Hasil Analisis Metode *K-Means* pada Data Sepeda Motor

Pada gambar 2 dapat dilihat bahawa berdasarkan hasil dari 15 kali pengujian di atas didapat nilai rata-rata hasil pengujian dari *Precision* 76%, *recall* 76%, dan *accuracy* 81%. Adapun nilai *precision* tertinggi adalah pada pengujian ke-1 dan ke-2 dengan nilai 87%. Nilai *recall* tertinggi adalah pada pengujian ke- 1,2 dan 3 senilai 85%. Nilai akurasi tertinggi adalah pada pengujian ke-1 dan 2 senilai 88%.

4 Conclusion

Dari hasil 15 kali pengujian yang dilakukan dengan menggunakan 3 cluster data pengujian yang berbeda setiap pengujiannya, iterasi paling sedikit berhenti di iterasi ke-4 pada pengujian 1 dengan data uji 2, 87, 100 dan pada pengujian 2 dengan data uji 1, 6, 9. Dengan iterasi yang paling banyak berhenti di iterasi ke-9 pada pengujian ke-5 dengan data uji 57, 66, 84. Hasil analisis performansi k-means dari 15 pengujian dari setiap uji coba yang dilakukan, diperoleh nilai rata-rata Presicion sebesar 76%, nilai Recall sebesar 76% dan Accuracy sebesar 81%.

Saran terhadap penelitian ini adalah penerapan k-means untuk clustering ini masih bisa dikembangkan dengan menggunakan data lain yang lebih banyak lagi. Selain itu pengembangan juga masih dapat dilakukan

dengan menggabungkan K-means dengan metode seleksi fitur untuk lebih meningkatkan performa hasil clustering.

References

[1] L. Surtiningsih, A. T. Putri, N. A. Putra, R. Arniantya, D. A. Prabowo, and I. Cholissodin, "Klasifikasi Sepeda Motor Berdasarkan Karakteristik Konsumen Dengan Metode K-Nearest Neighbour Pada Big Data Menggunakan Hadoop Single Node Cluster," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 2, p. 81, 2017.

[2] G. Abdillah, F. A. Putra, and F. Renaldi, "Penerapan Data Mining Pemakaian Air Pelanggan Untuk Menentukan Klasifikasi Potensi Pemakaian Air Pelanggan Baru Di Pdam Tirta Raharja Menggunakan Algoritma K-Means," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Data Mining, pp. 18–19, 2016.

[3] G. Abdurrahman, "Clustering Data Ujian Tengah Semester (UTS) Data Mining," *J. Sist. Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 71–79, 2016.

[4] W. Wardhani, Anindya Khrisna, "Implementasi Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Kajen Pekalongan," *J. Transform.*, vol. 14, no. 1, pp. 30–37, 2016.

[5] Nur, Fauziah; Zarlis, M.; Nasution, Benny Benyamin. "Penerapan Algoritma K-Means pada Siswa Baru Sekolah Menengah Kejuruan Untuk Clustering Jurusan". *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan*, 1.2: 100-105, 2017

[6] Windha Mega Pradnya Dhuhita, "Clustering Menggunakan Metode K-Means untuk Menentukan Status Gizi Balita," *J. Inform.*, vol. 15, no. 2, pp. 160–174, 2016.

[7] L. Maulida, "Penerapan Data Mining Dalam Mengelompokan Kunjungan Wisatawan Ke Objek Wisata Unggulan Di Prov. DKI Jakarta Dengan K-Means," *JISKa*, vol. 2, no. 3, pp. 167–174, 2018.

[8] S. Utara, "Implementasi Data Mining Untuk Memprediksi Pola Pembelian Sepeda Motor Pada Showroom CV . Viva Mas Motors Dengan Metode Algoritma C4 . 5," vol. 2, no. 2, pp. 34–38, 2018.

[9] B. M. Metisen and H. L. Sari, "Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila," vol. 11, no. 2, pp. 110–118, 2015.

[10] N. Rohmawati, S. Defiyanti, and M. Jajuli, "Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa," *J. Ilm. Teknol. Inf. Terap.*, vol. I, no. 2, pp. 62–68, 2015.

[11] I. Menarianti, "Klasifikasi data mining dalam menentukan pemberian kredit bagi nasabah koperasi," *J. Ilm. Teknosains*, vol. 1, no. 1, pp. 1–10, 2015.