

# PENENTUAN PRODUK YANG DIMINATI PASAR MENGUNAKAN ALGORITMA K-MEANS

**Muhammad Irwan Mahfud**

Sistem Informasi, Fakultas Teknologi Informasi, Universitas Hasyim Asy'ari

Email : [irwan.unhasy@gmail.com](mailto:irwan.unhasy@gmail.com)

**Achmad Imam Agung**

Sistem Informasi, Fakultas Teknologi Informasi, Universitas Hasyim Asy'ari

Email : [imamagung@unhasy.ac.id](mailto:imamagung@unhasy.ac.id)

**Indana Lazulfa**

Sistem Informasi, Fakultas Teknologi Informasi, Universitas Hasyim Asy'ari

Email : [indanazulf@gmail.com](mailto:indanazulf@gmail.com)

## Abstrak

Pengelompokan untuk mendapatkan kesamaan kelas serta membagi menjadi beberapa kelas adalah salah satu proses dari data mining. Keakuratan pengelompokan menjadi faktor penting untuk menentukan produk tersebut diminati pasar. Tujuan dari penelitian ini adalah menentukan produk yang tergolong paling diminati, diminati dan kurang diminati pasar, sehingga membantu dalam pengambilan keputusan. Penelitian ini menggunakan metode CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah metodologi dari data mining yang digunakan untuk melakukan analisis masalah pada proses bisnis atau unit penelitian. Algoritma K-Means digunakan untuk pengelompokan produk yang diminati pasar. Algoritma K-Means mempartisi kesamaan kelas berdasarkan parameter yang telah ditentukan, dengan menghitung jarak *centroid* pada suatu kelas. Penelitian ini menghasilkan sebuah sistem informasi penentuan produk yang diminati pasar. Dari hasil pengujian menggunakan enam parameter yaitu, jumlah transaksi, volume penjualan, kategori produk, keragaman produk, rata-rata penjualan dan jumlah stok dengan data transaksi sebanyak 1.235 transaksi. Diperoleh tiga *cluster* yang terbaik, telah dilakukan uji performa menggunakan metode Elbow dengan selisih nilai SSE terbanyak sebesar 28.00782.

**Kata Kunci:** Data mining, K-Means, *clustering*, produk diminati pasar.

## Abstract

Grouping to get class similarity and dividing into several classes is one of the processes of data mining. The accuracy of the grouping is an important factor to determine the product's interest in the market. The purpose of this study is to determine which products are classified as the most desirable, desirable and less desirable markets, so that petrified in decision making. This research uses the CRISP-DM (*Cross Industry Standard Process for Data Mining*) method is a methodology from minimum data that is used to analyze problems in business processes or research units. K-Means algorithm is used for grouping products that are of interest to the market. K-Means algorithm partitioned class similarity based on predetermined parameters, by calculating the centroid distance in a class. This research resulted in a product determination information system that is of interest to the market. From the test results using six parameters, namely, the number of transactions, sales volume, product categories, product diversity, average sales and number of stocks with transaction data of 1,235 transactions. Obtained the three best clusters, performance testing has been done using the Elbow method with the most SSE difference of 28,00782.

**Keywords:** Data mining, K-Means, *clustering*, products market demand.

## PENDAHULUAN

Pelaku UMKM (Usaha Micro Kecil Menengah) di Indonesia semakin berkembang seiring berjalannya waktu, ada banyak jenis usaha yang dijalankan seperti, penjualan baju, kuliner, konsultasi belajar, penyedia belanja online dan lain-lain. Vaska Store merupakan UMKM yang bergerak dibidang penjualan baju bertempat di Kabupaten Tulungagung. Produk yang dijual meliputi baju gamis, tunik, piyama wanita dan lain-lain. Pada saat ini Vaska Store belum pernah melakukan analisis data penjualan sehingga banyak produk menumpuk karena kurang diminati dan ada produk yang kosong ketika banyak diminati pembeli. Maka dari itu, penulis akan menggunakan data mining untuk penggalian data dengan melakukan pengelompokan menggunakan algoritma K-Means. K-Means bertujuan melakukan pengelompokan di kelas yang tepat pada data penjualan (Fintri dan Eni, 2019:110).

Dalam melakukan penggalian data diperlukan histori data penjualan, teknik data mining diperlukan apabila data penjualan banyak dan besar. Pada penelitian ini menggunakan algoritma K-Means untuk menentukan produk yang tergolong data produk yang paling diminati, diminati dan kurang diminati. Produk akan dikelompokkan menjadi beberapa *cluster* yang dibedakan berdasarkan parameter yang telah ditentukan. Parameter untuk melakukan *clustering* ada enam, diantaranya: kategori produk, ragam produk, jumlah stok, jumlah transaksi, volume penjualan dan rata-rata penjualan.

Algoritma K-Means merupakan salah satu algoritma *partition clustering* yang bertujuan untuk mengelompokkan data berdasarkan karakteristik yang mirip, jika suatu data memiliki karakteristik yang sama maka akan ditempatkan dalam satu *cluster* jika karakteristiknya berbeda akan ditempatkan dalam *cluster* yang lain (Mahmudi dkk, 2020:21).

Hasil pengelompokan sangat bergantung pada jumlah  $k$  yang ditentukan. Metode K-Means adalah teknik pengelompokan sederhana dan sangat cepat untuk menentukan jumlah *cluster* yang tepat dalam kumpulan data. Beberapa optimasi mengusulkan metode untuk menentukan jumlah  $k$  yang terbaik, salah satunya adalah metode Elbow. Metode Elbow sangat mudah di implementasikan, dengan melihat nilai SSE (*Sum of Square Error*) yang berbentuk siku pada grafik Elbow (Syakur dkk, 2018:2).

Sebelumnya sudah pernah dilakukan penelitian tentang *clustering* data bisnis ritel pakaian yang memiliki banyak cabang di Akwa Ibom Nigeria. Tujuan dari penelitian ini sebagai perencanaan strategi pelayanan pelanggan, identifikasi produk yang cocok untuk setiap pelanggannya dan memprediksi pelanggan yang curang. Parameter yang digunakan adalah jumlah rata-rata barang yang dibeli oleh pelanggan perbulan dan jumlah rata-rata kunjungan pelanggan per bulan. Algoritma ini memiliki ukuran kemurnian 0,95 yang menunjukkan 95% segmentasi pelanggan yang akurat. (Chinedu dkk, 2015). Berdasarkan penelitian diatas dapat disimpulkan K-Means memiliki hasil yang akurat, K-Means lebih optimal karena data yang digunakan sudah berbentuk angka jadi langsung dapat diolah dengan algoritma K-Means.

Sistem informasi penentuan produk yang diminati pasar diperlukan untuk membantu penelitian ini yang digunakan untuk mengolah data transaksi penjualan toko. Sistem ini diharapkan memberikan kemudahan terhadap bagian manajemen Vaska Store karena sistem analisa data berbasis web akan membantu menentukan strategi promosi produk yang kurang diminati dan manajemen stok yang efisien.

## METODE

Penelitian ini menggunakan tahapan metodologi data mining *Cross Industry Standard Process for Data Mining* (CRISP-DM). CRISP-DM adalah metodologi yang sering digunakan oleh para ahli dalam analisis penambangan data (*data mining*) sebagai dasar pemecahan masalah dari proses bisnis atau unit penelitian. Tahapan penelitian ini mengacu pada enam tahap, yaitu: (1) Pemahaman Bisnis (*Bussiness Understanding*), merupakan tahapan tentang pemahaman substansi dari proses data mining di perusahaan, dilihat dari kebutuhan bisnis. (2) Pemahaman Data (*Data Understanding*), merupakan tahapan pengumpulan data awal, mempelajari data yang tepat untuk proses data mining. Fase ini mencoba memahami masalah yang berkaitan dengan kualitas data, mencari *subset* data yang menarik. (3) Persiapan Data (*Data Preparation*), tahap ini merupakan fase padat karya. Kegiatan yang dilakukan antara lain mentransformasikan data kedalam database untuk bahan data mining. (4) Pemodelan (*Modeling*) merupakan tahapan dalam memilih

teknik data mining yang akan digunakan, algoritma yang akan digunakan, menentukan atribut yang terbaik untuk digunakan. (5) Evaluasi (*Evaluation*) tahap interpretasi dari hasil data mining, ditujukan dalam proses pemodelan agar sesuai dengan tujuan yang ingin dicapai. (6) Penyebaran (*Deployment*), merupakan tahap presentasi dari hasil yang didapat, bisa berupa penyusunan laporan atau visualisasi hasil pengetahuan yang didapat (Adiana dkk, 2018:25).

Penelitian ini menggunakan data yang didapat dari Vaska Store Tulungagung. Data yang dikumpulkan berupa data transaksi penjualan, data laporan penjualan dan data laporan stok. Data tersebut kemudian dilakukan persiapan data, diantaranya adalah seleksi data untuk pengambilan parameter yang dibutuhkan dan tranformasi data untuk mengubah parameter yang berupa data kategorik menjadi numerik.

Tahapan yang dilakukan untuk mempersiapkan data diantaranya memilih parameter yang akan dianalisa, menyiapkan data mentah, tranformasi data hingga siap untuk proses *clustering*.

Tabel 1. Hasil Seleksi Data

Data Mentah	Variabel
Data Transaksi Penjualan	Jumlah Transaksi
Laporan Penjualan	Kategori Produk
	Keragaman Produk
	Volume Penjualan
Laporan Stok Barang	Jumlah Stok
Jumlah Transaksi dibagi volume penjualan	Rata-rata penjualan

Tabel 2. Tranformasi Kategori Produk

Kategori Produk	Jumlah Transaksi	Transformasi
Gamis	1007	6
Tunik	441	5
Kaftan	232	4
Piyama	218	3
Overall	207	2
Jaket	111	1

Tabel 3. Tranformasi Keragaman Produk

Keragaman Produk	Jumlah Barang	Transformasi
Gamis	82	6
Tunik	36	5
Kaftan	24	4
Overall	14	3
Jaket	14	2
Piyama	13	1

Tabel 4. Data Siap *Clustering*

Nama Produk	Jumlah Transaksi	Volume Penjualan	Kategori Produk	Keragaman Produk	Rata-rata Penjualan	Stok
Akazam Overall	31	54	2	3	1.74	31
Alma Tunik	59	232	5	5	3.93	19
Amarilia Kaftan	34	83	4	4	2.44	23
Angela Overall	45	68	2	3	1.51	31
Asymmetric Kaftan	27	54	4	4	2.00	45

Adapun teknik analisa yang dipakai dalam penelitian ini dengan menggunakan algoritma K-Means. Algoritma ini termasuk dalam data mining *partitioning clustering*. Proses *clustering* dilakukan untuk mengelompokkan produk yang paling laris berdasarkan parameter yang telah ditentukan.

Tahapan melakukan *clustering* atau pengelompokan dengan algoritma K-Means adalah sebagai berikut :

1. Tentukan banyaknya *cluster k*.
2. Penentuan *k* titik *cluster* ini dilakukan dengan bermacam cara. Namun paling sering dilakukan adalah secara random.
3. Menghitung jarak data dengan titik centroid menggunakan rumus *euclidean* (persamaan 1)

$$d(x_i, \mu_j) = \sqrt{\sum (x_i - \mu_j)^2} \quad (1)$$

Dalam persamaan 1,  $d(x_i, \mu_j)$  merupakan titik dokumen,  $x_i$  adalah data awal jarak dari titik centroid, sedangkan  $\mu_j$  adalah titik centroid pada *cluster ke-j*.

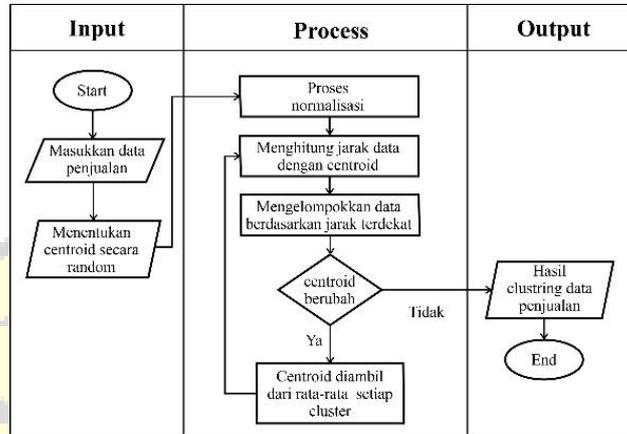
4. Kelompokkan data berdasarkan jarak paling dekat dengan centroid, kemudian perbaharui nilai centroid baru dengan menghitung rata-rata data *cluster* yang terbentuk (persamaan 2)

$$\mu_j^{(t+1)} = \frac{1}{N_{sj}} \sum_{j \in s_j} x_j \quad (2)$$

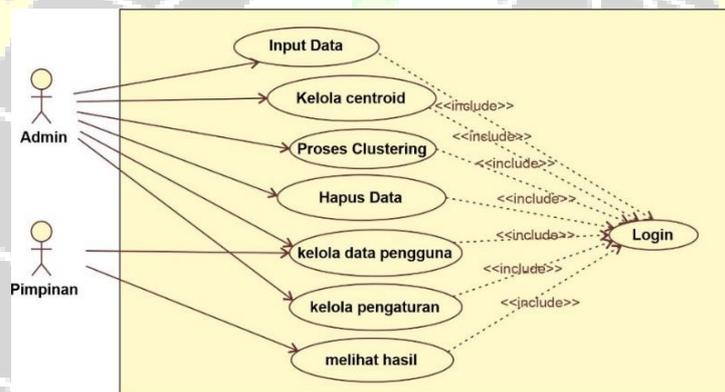
Dalam persamaan 2,  $\mu_j^{(t+1)}$  merupakan centroid baru pada iterasi ke (t+1) dan  $N_{sj}$  adalah banyaknya data pada *cluster sj*.

- Lakukan iterasi dengan centroid yang baru. Sampai pusat *cluster* tidak mengalami perubahan maka proses *clustering* selesai (Saut dkk, 2019:68).

Rancang bangun sistem dilakukan sebagai langkah awal pembuatan sistem informasi penentuan produk yang diminati pasar. Tahap pertama perancangan adalah analisa kebutuhan sistem dilanjutkan dengan pembuatan *flowchart* sistem dan pembuatan model UML (*Unified Modeling Language*).



Gambar 1. *Flowchart* Sistem Penentuan Produk Diminati Pasar

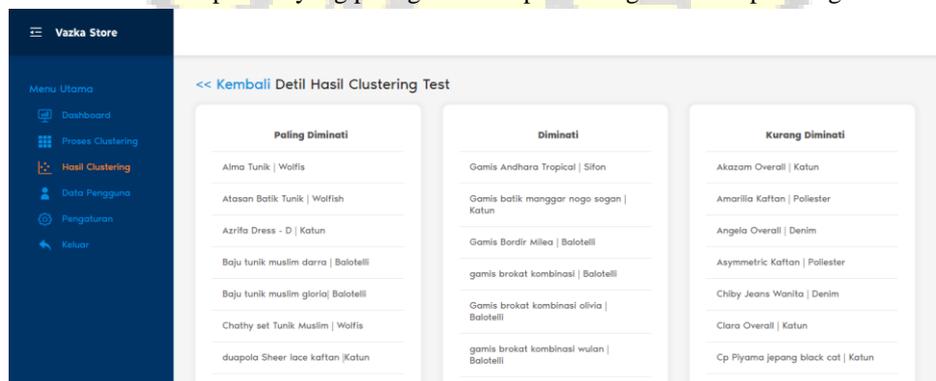


Gambar 2. *Usecase* Sistem Penentuan Produk Diminati Pasar

## HASIL DAN PEMBAHASAN

### Hasil

Sistem informasi penentuan produk yang diminati pasar ini memiliki sebuah *database* atau media penyimpanan yang berfungsi menyimpan data yang dibutuhkan oleh sistem. Database yang digunakan adalah MySQL. Sistem diimplementasikan dengan bahasa pemrograman PHP. Sistem yang dibangun bertujuan dalam menentukan produk yang paling diminati pasar dengan menerapkan algoritma K-Means.



Gambar 3. Tampilan Hasil Penentuan Produk Diminati Pasar

## Pembahasan

Algoritma K-Means akan mengelompokkan anggota berdasarkan ukuran jarak tertentu (persamaan 1). Dalam melakukan pengelompokan K-Means diperlukan data penjualan, data penjualan ini didapat dengan cara meng-import data siap pada sistem.

a) Menentukan Kelompok (*Cluster*)

Tentukan banyaknya *cluster* yang akan dibentuk dan menentukan titik *cluster* pertama. Penentuan pusat titik centroid pertama ditentukan secara acak (*random*) seperti pada Tabel 5. berikut:

Tabel 5. Titik Centroid Pertama

Pusat Cluster	Jumlah Transaksi	Volume Penjualan	Kategori Produk	Keragaman Produk	Rata-rata penjualan	Stok
1	59	232	5	5	3.93	19
2	18	65	6	6	3.61	70
3	2	3	4	4	1.50	79

b) Menentukan jarak data pada setiap *cluster*

Setelah menentukan nilai titik pusat setiap *cluster*, proses selanjutnya adalah menentukan jarak data ke setiap pusat *cluster*.

Jarak data ke-1 ( $x_1$ ) dengan *centroid* ( $\mu_1$ ) pada *Cluster* 1 adalah:

$$d(x_1, \mu_1) = \sqrt{(1.511 - 1.860)^2 + (1.712 - 4)^2 + (1.2 - 1.8)^2 + (1.4 - 1.8)^2 + (1.078 - 1.366)^2 + (1.689 - 1.827)^2} = 2.444$$

Jarak data ke-1 ( $x_1$ ) dengan *centroid* ( $\mu_2$ ) pada *Cluster* 2 adalah:

$$d(x_1, \mu_2) = \sqrt{(1.511 - 1.255)^2 + (1.712 - 1.543)^2 + (1.2 - 2)^2 + (1.4 - 2)^2 + (1.078 - 1.169)^2 + (1.689 - 1.241)^2} = 1.141$$

Jarak data ke-1 ( $x_1$ ) dengan *centroid* ( $\mu_3$ ) pada *Cluster* 3 adalah:

$$d(x_1, \mu_3) = \sqrt{(1.511 - 1)^2 + (1.712 - 1)^2 + (1.2 - 1.6)^2 + (1.4 - 1.6)^2 + (1.078 - 1)^2 + (1.689 - 1.137)^2} = 1.131$$

Berikut hasil dari penghitungan iterasi pertama dari jarak data ke setiap titik *cluster*. Titik *cluster* di terdekat akan dikelompokkan dengan titik *cluster* yang sama.

Tabel 6. Titik Iterasi Pertama

Data	C1	C2	C3	Jarak Terdekat
1	2.444612	1.14156	1.13138	3
2	0	2.619364	3.229597	1
3	2.298046	0.842309	1.111057	2
4	2.486965	1.187643	1.179159	3
5	2.647336	0.662166	0.704641	2
6	0.134098575	2.525273059	3.134394855	1
7	1.877984393	0.980232977	1.402988102	2
8	1.33246594	1.480305445	2.065785023	1
9	1.65970039	1.228876187	1.800326448	2
10	0.287485614	2.483649682	3.088491967	1
11	2.963813	1.067089	0.628571	3
12	2.527836	1.089087	1.01057	3
13	2.695243	1.232966	1.03162	3
14	2.796277	1.218881	0.942113	3
15	2.6012	1.249467	1.108448	3
16	2.06122	0.998168	1.350343	2
...	...	...	...	...
185	3.155567	0.806301	0.150315	3

c) Tugaskan lagi dengan titik *cluster* terbaru

Selanjutnya dapat dilakukan proses penentuan pusat titik centroid kedua untuk melakukan proses iterasi selanjutnya. Titik centroid ini didapatkan dengan cara menghitung rata-rata data pada setiap *cluster*.

Tabel 7. Titik Centroid Kedua

Pusat Cluster	Jumlah Transaksi	Volume Penjualan	Kategori Produk	Keragaman Produk	Rata-rata Penjualan	Stok
1	1.658	3.410	1.916	1.916	1.428	1.885
2	1.170	1.454	1.942	1.942	1.249	1.398
3	1.101	1.144	1.520	1.529	1.081	1.273

Setelah dilakukan beberapa kali proses iterasi maka dapat diketahui dari hasil anggota pada setiap cluster. Pada proses ini terdapat sembilan kali proses iterasi sampai pusat titik centroid tidak berubah. Berikut Tabel 8. adalah hasil akhir dari keanggotaan dari setiap cluster.

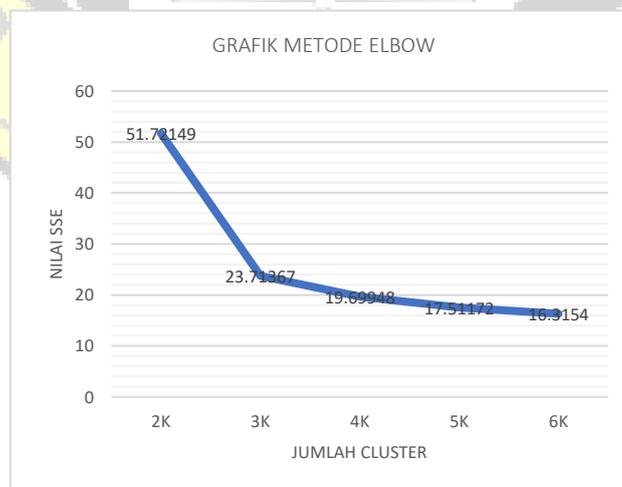
Tabel 8. Hasil Clustering

Paling Diminati	Diminati	Kurang Diminati
Alma Tunik   Wolfis	Gamis Andhara Tropical   Sifon	Akazam Overall   Katun
Atasan Batik Tunik   Wolfish	Gamis batik manggar nogo sogan   Katun	Amarilia Kaftan   Poliester
Azrifa Dress - D   Katun	Gamis Bordir Milea   Balotelli	Angela Overall   Denim
Baju tunik muslim darra   Balotelli	gamis brokat kombinasi   Balotelli	Asymmetric Kaftan   Poliester
Baju tunik muslim gloria   Balotelli	Gamis brokat kombinasi olivia   Balotelli	Chiby Jeans Wanita   Denim

Pada pengujian menggunakan metode Elbow menampilkan grafik yang berisi nilai SSE disetiap percobaan jumlah cluster antara 2 hingga 6. Gambar 4. merupakan grafik metode Elbow yang menunjukkan bentuk siku pada cluster ke tiga dengan nilai SSE sebesar 23.71367, karena semakin kecil nilai SSE maka semakin besar jumlah cluster yang ditentukan (Syakur dkk, 2018:5). Penentuan produk yang paling diminati pasar diambil dari rata-rata terbesar setiap cluster, apabila didalam suatu cluster memiliki rata-rata paling besar maka produk tersebut dinyatakan paling diminati sebaliknya jika didalam suatu cluster memiliki rata-rata terkecil maka produk tersebut dinyatakan kurang diminati. Dengan demikian hasil selisih yang paling besar pada 3K ditunjukkan pada Table 9. dibawah ini:

Tabel 9. Nilai SSE cluster 2 sampai 6

Cluster	Hasil SSE	Selisih
2K	51.72149	0
3K	23.71367	28.00782
4K	19.69948	4.01419
5K	17.51172	2.18776
6K	16.3154	1.19632



Gambar 4. Grafik hasil uji Metode Elbow

Dari nilai SSE pada Tabel 9. dapat diketahui hasil *cluster* terbaik adalah pada *cluster* 3K, dimana jumlah pada *cluster* satu adalah 38 produk dengan nilai rata-rata jumlah transaksi 27 kali, jumlah volume penjualan 122 dan jumlah stok yang ada sebanyak 21. *Cluster* dua ada 43 produk dengan nilai rata-rata jumlah transaksi 13, jumlah volume penjualan 25 dan jumlah stok yang ada sebanyak 61. *Cluster* tiga memiliki 104 produk dengan nilai rata-rata jumlah transaksi 6 kali, jumlah volume penjualan 16 dan jumlah stok yang ada sebanyak 72.

## PENUTUP

### Simpulan

1. Pembuatan sistem informasi penentuan produk diminati pasar dimulai dengan tahapan analisis kebutuhan sistem, pembuatan *flowchart* sistem dan pembuatan model UML (*Unified Modelling Language*) dari pembuatan model UML selanjutnya di implementasikan ke dalam bahasa pemrograman PHP. Proses pengujian sistem menggunakan *blackbox testing*, dari hasil pengujian menunjukkan semua modul dapat berjalan dengan baik sesuai fungsinya.
2. Algoritma K-Means digunakan sebagai teknik pengelompokan dalam sistem informasi penentuan produk yang diminati pasar. Proses *clustering* ini menggunakan enam parameter, yaitu: Jumlah Transaksi, Volume Penjualan, Kategori Produk, Keragaman Produk, Rata-rata Penjualan dan Jumlah Stok. Proses normalisasi dilakukan untuk membalik nilai dari jumlah stok, selanjutnya menentukan titik centroid dan akan dihitung jarak *euclidean* pada setiap data. Proses iterasi dilanjutkan sampai jarak centroid tidak berubah. Penentuan produk yang paling diminati didapatkan dari hasil rata-rata terbesar pada setiap *cluster*, apabila suatu *cluster* mempunyai rata-rata terbesar maka produk tersebut dinyatakan paling diminati. Pada pengujian Elbow didapatkan jumlah *cluster* terbaik adalah 3 dengan selisih nilai SSE terbanyak sebesar 28.00782.

### Saran

1. Data transaksi penjualan yang digunakan sebaiknya dengan rentang waktu lebih dari satu tahun.
2. Penambahan parameter juga dapat dilakukan untuk pengembangan penelitian. Contoh parameter yang dapat digunakan yaitu warna produk, ukuran produk dan lain-lain.
3. Penelitian ini juga dapat dikembangkan dengan menggabungkan metode lainnya untuk membantu mencari nilai *k* yang lebih optimal.

### DAFTAR PUSTAKA

- Adiana, B. E., Soesanti I., dan Permanasari, A. E., 2018. "Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering". *JUTEI*, Vol. 2, No.1, Hal. 23-32.
- Ezenkwu, C. P., Ozuomba, S., dan Kalu, C., 2015. "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services". *International Journal of Advanced Research in Artificial Intelligence*, Vol. 4, No.10, Hal. 40-44.
- Indriyani, F., dan Irfiani E., 2019. "Clustering Data Penjualan pada Toko Perlengkapan Outdoor Menggunakan Metode K-Means". *JUITA: Jurnal Informatika*, Vol. 7, No. 2, Hal. 109-113.
- Mahmudi, I., Indriyanti, A. D., dan Lazulfa, I., 2020. "Penerapan Algoritma K-Means *clustering* sebagai strategi promosi penerimaan mahasiswa baru pada Universitas Hasyim Asy'ari Jombang". *INOVATE*, Vol. 4 No. 2 Hal. 20-27.
- Mardalius, 2018. "Pengelompokan Data Penjualan Aksesoris Menggunakan Algoritma K-Means". *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, Vol. IV, No. 2, Hal. 401– 411.
- Syakur M. A., Khotimah, B. K., Rochman, E. M., dan Satoto, B. D., 2018. "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster". *IOP Conference Series: Materials Science and Engineering* Vol. 336, Surabaya, 9 November 2017, Hal. 1-6.
- Tamba, S. P., Kesuma, F. T., dan Feryanto, 2019. "Penerapan Data Mining untuk Menentukan Penjualan Sparepart Toyota Dengan Metode K-Means Clustering". *Jurnal Sistem Informasi Ilmu Komputer Prima*, Vol. 2, No. 2, Hal. 67-72.