PENGGUNAAN ALGORITMA K-MEANS UNTUK MENGANALISA PENJUALAN DI BIGMART

Bayu Rimbi Asmoro[1], Agus Wibowo[2], Ari Fiqri Aryadi[3]
[1,2,3)]Magister Ilmu Komputer, Universitas Budi Luhur
[1]2111600124@student.budiluhur.ac.id, [2]2111600736@student.budiluhur.ac.id,
[3]2111600561@student.budiluhur.ac.id

Abstrak

BigMart mencatatkan data terkait penjualan produk untuk diolah sebagai bahan pengambilan keputusan untuk pembelian produk sebagai dasar stok dan analisis sales di masa depan. BigMart dapat memahami karakteristik produk dan berbagai jenis outlet yang tersebar untuk memainkan peranan sangat penting dalam meningkatkan penjualan. Penjualan tidak mungkin dipisahkan dari tipe-tipe kategori barang yang laku dan tidak laku. Untuk itu perlu analisis yang mendalam untuk mengetahui sejauh mana data barang yang sangat laku, laku, dan kurang laku, untuk merencanakan pembelian kembali setelah barang kosong di gudang. Perusahaan ini perlu mengelompokkan ke berbagai macam cluster untuk mengetahui barang mana yang diminati dan kurang diminati konsumen. Kami menggunakan dataset ini dari Kaggle. Tools yang kami gunakan adalah algoritma k-means, tahapan preprocessing menggunakan Transformasi Z, dan software Rapidminer. Hasil perhitungan menggunakan rumus di Microsoft Excel terdapat 4 kategori barang paling laku, 6 kategori barang laku, dan 6 kategori barang kurang laku dengan terlebih dahulu mengelompokkan titik pusat antar cluster, dan dibagi 3 titik acak C_0, C_1, dan C_2. Pada hasil perhitungan dengan software Rapidminer, hasil Davies Bouldin Index pada performance vector menunjukkan angka 0,497, dari angka ini hasil yang didapat sudah cukup baik, karena berada pada angka>=0.

Kata kunci—, BigMart, *Clustering, K-Means, Rapidminer*

1. **Pendahuluan**

Penjualan di toko swalayan BigMart yang disajikan datanya di *Kaggle.com* dapat dianalisis dari total penjualan berbagai macam kategori yang disajikan. Banyaknya data sales yang bisa ditarik berdasarkan variabel dapat ditentukan jenis barang mana yang laku keras, laku, dan kurang laku. BigMart didirikan tahun 2018 berlokasi di Tangerang Selatan. Tipe bisnis dari BigMart adalah distributor kelas menengah dengan tagline "Produk Sembako dan Grosiran."

BigMart memerlukan rekomendasi barang mana yang perlu mereka stok dalam jumlah tertentu. Masalah yang sering dijumpai yaitu ada barang yang kurang diminati konsumen hingga banyak barang menumpuk dan tidak terjual, padahal sudah dibeli dalam jumlah besar. Pada akhirnya mereka harus menjual produk tersebut dengan memberikan harga promosi agar tidak mengalami kerugian yang besar.

Penelitian dilakukan dengan menggunakan *k-means* dan membuat *clustering* antara barang paling laku, laku dan kurang laku (Wahyudi et al. 2020). Data sementara dari tabel *pivot* menunjukkan tipe buah dan sayuran (*fruit and vegetables*) adalah angka yang dua paling tinggi. Didapatkan hasil penelitian setelah berulang kali dicoba sampai iterasi ke-5, titik pusat tidak berubah dan tetap antar *cluster*.

Hasil pengelompokan dibagi menjadi tiga *cluster*. *Cluster* pertama merupakan barang paling laku (C_0) sebanyak 4 kategori barang, barang yang laku (C_1) sebanyak 6 kategori barang, dan yang terakhir barang kurang laku (C_2) sebanyak 6 kategori barang.

Pada penelitian ke-5 penerapan metode *cluster*ing bisa menentukan pembelian dari stok barang yang memang dibutuhkan dengan segera, dari riset yang dilakukan didapatkan hasil bahwa kelompok barang yang laris adalah 4 item, sehingga pembelian stok barang diutamakan pada 4 item tersebut. Hasil penelitian iterasi ke-5 tersebut didapatkan kesimpulan.

Tujuan dari penelitian menggunakan *k-means* ini adalah untuk mengetahui *cluster* dari produk-produk yang dijual di BigMart secara iterative (Novia et al, 2020), sehingga data yang didapatkan dapat dijadikan sebagai rekomendasi bagi manajemen dalam merencanakan stok produk agar toko tidak mengecewakan pelanggan karena barang yang ingin dibeli tidak tersedia. Dari latar belakang yang telah disediakan sebelumnya, maka dapat dirumuskan permasalahan pada penelitian ini adalah bagaimana mengimplementasikan metode *k-means* dalam *clustering* produk penjualan pada BigMart dengan *cluster* produk tersebut kedalam tiga kelompok yaitu yang paling laku, laku, dan kurang laku.

2. **Kajian Pustaka**

a. Data Mining

Data Mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di database yang besar (Siregar & Pusphabuana, 2020). Data yang dibuat sangat bermacam-macam, hampir di semua sektor, seperti data di sektor bisnis atau perdagangan, sektor pendidikan, sektor keuangan dan sektor lainnya. Data yang bermacam-macam ini kemudian dikembangkan sehingga memberikan pemahaman yang aktual. *Data mining* merupakan pengamatan pada data dalam mendapatkan ikatan yang terbukti serta merumuskan yang sebelumnya tidak diketahui dengan cara terbaru disimpulkan dan bermanfaat untuk pemilik data tersebut.

Secara garis besar, *data mining* dapat dibagi menjadi dua kelompok utama, yaitu:

1) *Descriptive Mining*, yakni metode untuk menemukan ciri yang penting dalam suatu data. Ada beberapa teknik dalam diantaranya adalah*clustering*, *association*, dan *sequential mining*.
2) *Predictive Mining*, cara mencari metode untuk mendapatkan model metode pada masa yang akan datang.Contohnya adalah metode klasifikasi.

b. Clustering

Clustering merupakan suatu proses pengelompokan *record*, *observasi*, atau mengelompokan kelas yang memiliki kesamaan objek. *Clustering* sering digunakan sebagai tahap pertama dalam metode data mining (penambangan data). Beberapa tipe clustering antara lain*K-Means, Improved K-Means, Fuzzy C-Means, DBSCAN, K-Medoids (PAM), CLARANS*, dan *Fuzzy Subtractive*.

Clustering bertujuan meminimalkan fungsi objek kedalam satu kesatuan set *clustering* dari beberapa keragaman antar cluster. Analisis pengelompokan atau biasa disebut *clustering* adalah cara memisahkan data dalam suatu gabungan ke dalam beberapa grup yang datanya sama dalam suatu kelompok lebih besar daripada kesamaan data tersebut dengan data dalam kelompok lain.

Potensi *clustering* yaitu bisa dimanfaatkan lebih lanjut dalam berbagai aplikasi secara umum seperti pengklasifikasian, pengerjaan gambar, dan pemahaman

pola. Penerapan metode *clustering* mendapatkan pemahaman berupa penetapan beberapa *cluster* catatan data yang mempunyai kemiripan atribut.

*Clustering*digunakan untuk memprediksi dan analisa masalah bisnis. Untuk mememtakan segmen pasar, kelompok pelanggan, zona-zona wilayah tertentu, dan identifikasi objek.

c. Metode *K-Means*

K-Means adalah algoritma *clustering* yang masuk dalam kelompok *unspervised learning*, menggunakan sistem partisi yang membagi data menjadi beberapa bagian (Bungin, 2017). Metode ini untuk meminimalkan fungsi object dalam clustering yang bermaksud memniminalisasi perbedaan antar data dalam satu *cluster*.

Kelebihan:

1. Sangat mudah untuk diterapkan dan dipakai oleh peneliti.
2. Proses pengerjaan tidak memakan waktu yang lama.
3. Dapat digunakan dengan berbagai macam dataset.
4. Sudah banyak yang menggunakan, terkenal di kalangan peneliti.

Kekurangan:

1. K buah titik di inisialisasi secara *random* sehingga pengelompokan data yang dihasilkan dapat berbeda-beda dan jika di inisialisasi kurang baik, beberapa *cluster* yang dikelompokkan menjadi kurang maksimal.
2. Mengakibatkan*curse of dimensionality*. Jika data pelatihan memiliki dimensi yang sangat tinggi maka akan menjadi sulit.
3. Sulit menghitung dan mencari titik terdekat dengan K titik jika terdapat banyak sekali titik data.

Dalam metode *k-means*, kita perlu menentukan jumlah *cluster* k. Pada titik itu, fokus pengelompokan k dipilih secara acak.

Dasar algoritma *k-means* adalah:

1) Menentukan besar k, dan jumlah *cluster* yang diperlukan
2) Menandakan nilai k menjadi titik patokan pusat secara acak
3) Menghitung jarak data ke masing-masing nilai *centroid* dengan rumus *Euclidean Distance*:

$$D(i,j) = \text{akar } (x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 \quad \dots (1)$$

4) Gabungkan setiap hasil yang nilainya terendah dan dikelompokkan
5) Menentukan posisi titik *centroid* baru (k)
6) Kembali ke langkah sebelumnya jika titik *centroid* yang lama tidak sama dengan sebelumnya.

Pengelompokan k-means digunakan sebagai metode pengelompokan dengan alasan bahwa algoritma *K-Means* dapat menangani sejumlah data set besar secara efektif. Algoritma *K-Means* adalah sejenis metode pengelompokan berdasarkan partisi. Inti dari

algoritma *K-Means* adalah memilih titik pusat k secara acak dan mempartisi data sesuai dengan jarak antara data dan titik tengah k.

3. Metode Penelitian

Metode penelitian menjelaskan langkah yang digunakan dalam melakukan implementasi *clusteringk-means* untuk mengelompokkan data penjualan barang menjadi tiga *cluster*. Tahapan metodologi penelitian ditunjukkan pada gambar 1.

Penelitian ini menggunakan jenis data kuantitatif, yaitu berupa angka atau nominal data yang dapat dihitung. Setiap penelitian kuantitatif dimulai dengan menjelaskan konsep penelitian yang digunakan (Wanto et al, 2020). Dalam penelitian ini menggunakan *dataset* historis penjualan dan merupakan jenis data kuantitatif karana berupa angka atau nominal yang dapat dihitung. Sumber data yang digunakan merupakan data sekunder karena data yang didapat tidak secara langsung dan melalui media perantara yang telah dicatat dan diperoleh oleh pihak lain.

Data yang dibutuhkan dalam penelitian ini adalah data penjualan periode sebelumnya, pengumpulan data dalam penelitian ini berupa data set yang didapat dari *website kaggle.com* yang kemudian dianalisis dan diterapkan untuk mendapatkan hasil kesimpulan penjualan BigMart.

Gambar 1. Tahapan metode penelitian

4. Hasil dan Pembahasan

Data penelitian ini didapat dari situs website *kaggle.com*, dimana data yang digunakan sebanyak 16 data. *Output* yang diinginkan adalah mendapatkan hasil 3 *cluster* yaitu data penjualan dari yang kurang laris (C_0), data penjualan yang laris (C_1), dan data penjualan yang paling laris (C_2). Variabel atau atribut yang dipakai untuk pengelompokan penjualan ini terdiri atas 16 kategori produk yang terjual:

Tabel 1. Jumlah produk yang terjual berdasarkan kategori

No.	Tipe Barang	Qty	Sales
1	Baking Goods	648	1,265,525
2	Breads	251	553,237
3	Breakfast	110	232,299
4	Canned	649	1,444,151
5	Dairy	682	1,522,594

6	Frozen Foods	856	1,825,735
7	Fruits and Vegetables	1232	2,820,060
8	Hard Drinks	214	457,793
9	Health and Hygiene	520	1,045,200
10	Household	910	2,055,494
11	Meat	425	917,56
12	Others	169	325,518
13	Seafood	64	148,868
14	Snack Foods	1200	2,732,786
15	Soft Drinks	445	892,898
16	Starchy Foods	148	351,401
		8,523	**18,591,125**

a. Tahapan *Preprocessing*

Tahap *Preprocessing*adalah untuk menerangkan proses sebelum dilaksanakan proses selanjutnya (Prasetyowati, 2017). Dikarenakan data yang didapat memiliki rentang yang sangat jauh, maka perlu dilakukan proses transformasi data. Metode transformasi data yang digunakan adalah transformasi Z.

Transformasi Z bisa dilakukan apabila diperlukan untuk memanipulasi bilangan deret dengan rentang yang terlalu untuk mendapatkan bilangan deret dengan rentang yang lebih dekat.

Urutan langkah melakukan proses transformasi data untuk mengetahui nilai Z adalah

1. Mencari tahu nilai rata-rata */mean* dari atribut qty dan *sales*.
2. Mencari tahu nilai deviasi standar dari atribut qty dan *sales*.
3. Mencari tahu nilai Z dengan menggunakan persamaan

$$z = \frac{x-\mu}{\sigma} \dots (2)$$

Dimana :

Z : Nilai Z

X : Nilai atribut datapoint

μ : Nilai rata-rata atribut yang sama

σ : Deviasi standar dari atribut yang sama

Berdasarkan persamaan diatas maka diketahui nilai *Z-transformation* pada setiap data yang ada pada table adalah sebagai berikut:

Tabel 2. Nilai rata-rata pada atribut QTY dan Sales

AVERAGE_QTY	AVERAGE_SALES

532,6875	1.161.945

Tabel 3. Deviasi Standar pada atribut QTY dan Sales

STDDEV_QTY	STDEV_SALES
374,9934611	853839,2185

Tabel 4. Nilai Z pada atribut QTY dan Sales

No.	Tipe Barang	Qty	Sales
1	Baking Goods	0,308	0,121
2	Breads	-0,751	-0,713
3	Breakfast	-1,127	-1,089
4	Canned	0,310	0,331
5	Dairy	0,398	0,422
6	Frozen Foods	0,862	0,777
7	Fruits and Vegetables	1,865	1,942
8	Hard Drinks	-0,850	-0,825
9	Health and Hygiene	-0,034	-0,137
10	Household	1,006	1,047
11	Meat	-0,287	-0,286
12	Others	-0,970	-0,980
13	Seafood	-1,250	-1,186
14	Snack Foods	1,780	1,840
15	Soft Drinks	-0,234	-0,315
16	Starchy Foods	-1,026	-0,949

b. Penerapan *K-Means Clustering*

1). Langkah awal dalam memutuskan total *cluster*yang ingin dibuat adalah menyusun sebanyak 3 *cluster*. Kemudian pilih pusat pertama dari semua *cluster*, dimana pusat pertama yang dipilih secara random ditunjukkan pada tabel 5:

Tabel 5. Titik *Centroid* Awal

	BARIS	QTY	SALES
C_0	5	0,398	0,422
C_1	13	-1,250	-1,186
C_2	9	-0,034	-0,137

2). Hitung semua atribut qty dan sales pada datapoint dengan menggunakan persamaan

$$d(a,b) = \sqrt{(xa - yb)^2 + (xa - yb)^2 + (xn - yn)^2} \ldots. (3)$$

Dimana

d(a,b)	: Jarak data ke (i) ke pusat *cluster* j
xa	: Nilai atribut data ke-i
yb	: Nilai atribut data pada pusat data

berdasarkan perhitungan diatas, maka diketahui jarak setiap data terhadap titik *cluster* adalah sebagai berikut:

Tabel 6 . Hasil iterasi ke 1 pada *dataset*

No.	**Tipe Barang**	**Cluster Iterasi 1**		**Kelompok**
		QTY	**SALES**	
1	Baking Goods	0.308	0.121	C_0
2	Breads	-0.751	-0.713	C_1
3	Breakfast	-1.127	-1.089	C_1
4	Canned	0.310	0.331	C_0
5	Dairy	0.398	0.422	C_0
6	Frozen Foods	0.862	0.777	C_0
7	Fruits and Vegetables	1.865	1.942	C_0
8	Hard Drinks	-0.850	-0.825	C_1
9	Health and Hygiene	-0.034	-0.137	C_2
10	Household	1.006	1.047	C_0
11	Meat	-0.287	-0.286	C_2
12	Others	-0.970	-0.980	C_1
13	Seafood	-1.250	-1.186	C_1
14	Snack Foods	1.780	1.840	C_0
15	Soft Drinks	-0.234	-0.315	C_2
16	Starchy Foods	-1.026	-0.949	C_1

Berdasarkan perhitungan pada iterasi ke 1 maka diketahui variabel WCV (*WithinClusterVariation* sebesar 10.767 didapat dari kuadrat minimal jarak ke pusat *cluster*. Setelah mengetahui nilai WCV peneliti kemudian mencari nilai BCV dengan menggunakan parameter jarak dari *cluster* C_0 ke C_1, C_0 ke C_2, dan C_1 ke C_2 sesuai dengan tabel 7 berikut:

Tabel 7. Jarak antar pusat *cluster*

Jarak Antar Pusat *Cluster*		
C_0	C_1	2,303
C_0	C_2	0,707
C_1	C_2	1,606
BCV		4,616

Setelah peneliti melakukan pengulangan perhitungan, ditemukan bahwa anggota *cluster* dan titik *centroid* tidak berubah setelah melakukan pengulangan ke 5. Berikut adalah tabelnya:

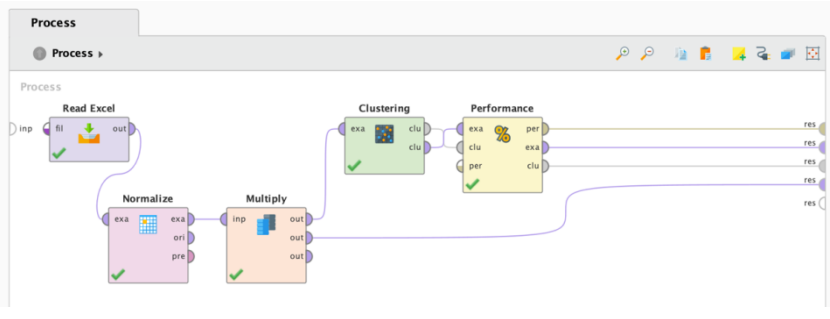
Tabel 8. Hasil iterasi ke 5 pada *dataset*

No.	**Tipe Barang**	**Cluster Iterasi 5**		**Kelompok**
		QTY	**SALES**	
1	Baking Goods	0.308	0.121	C_2
2	Breads	-0.751	-0.713	C_1
3	Breakfast	-1.127	-1.089	C_1
4	Canned	0.310	0.331	C_2
5	Dairy	0.398	0.422	C_2
6	Frozen Foods	0.862	0.777	C_0
7	Fruits and Vegetables	1.865	1.942	C_0
8	Hard Drinks	-0.850	-0.825	C_1
9	Health and Hygiene	-0.034	-0.137	C_2
10	Household	1.006	1.047	C_0
11	Meat	-0.287	-0.286	C_2
12	Others	-0.970	-0.980	C_1
13	Seafood	-1.250	-1.186	C_1
14	Snack Foods	1.780	1.840	C_0
15	Soft Drinks	-0.234	-0.315	C_2
16	Starchy Foods	-1.026	-0.949	C_1

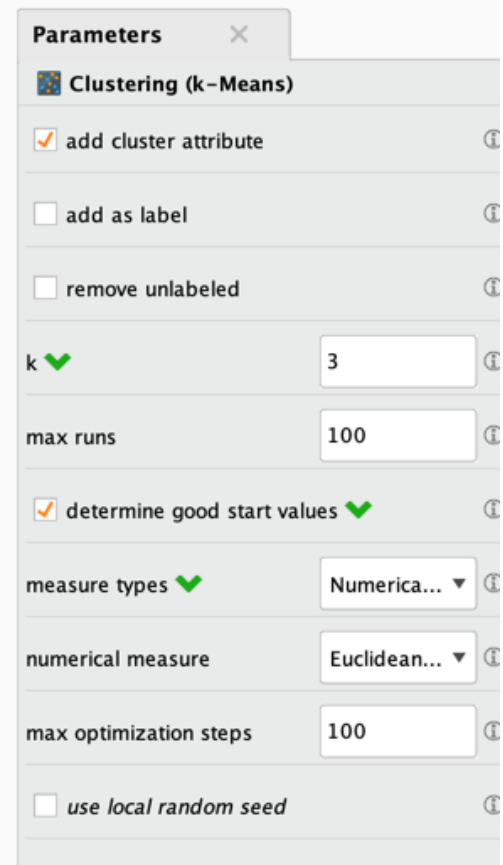
c. Implementasi*Tool Rapidminer*

Selain menggunakan beberapa algoritma, teknik evaluasi dapat juga menggunakan *software Rapidminer*, atau bahasa pemrograman *Python* (Jollyta et al, 2021).

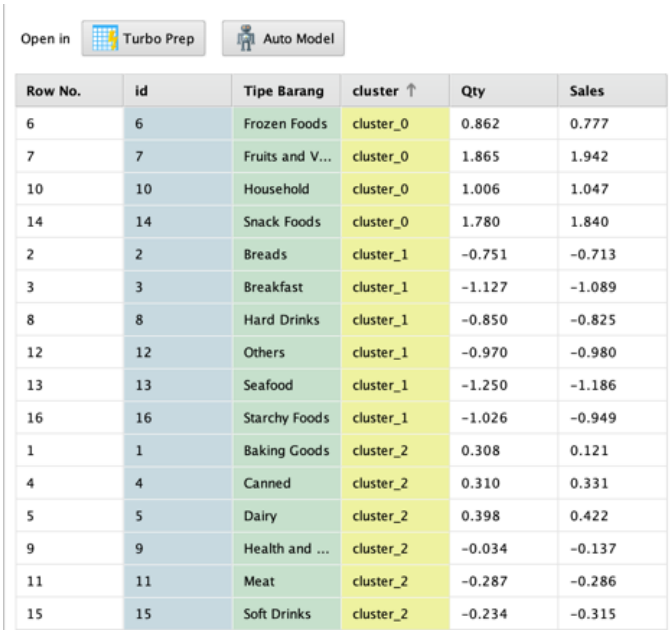
Pada tahap pengetesan algoritma ini, untuk menunjukkan bukti pada tahap analisis sebelumnya dan pengetesan secara manual, maka dari itu perlu dilakukan pengetesan lagi untuk mengkategorikan data untuk menaikkan jumlah penjualan menggunakan algoritma *k-means*.



Gambar 2. Proses menggunakan software *Rapidminer*



Gambar 3. Parameter *k-means* yang digunakan

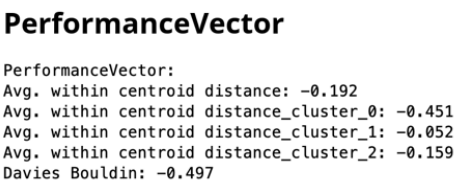


Open in Turbo Prep Auto Model

Row No.	id	Tipe Barang	cluster	Qty	Sales
6	6	Frozen Foods	cluster_0	0.862	0.777
7	7	Fruits and V...	cluster_0	1.865	1.942
10	10	Household	cluster_0	1.006	1.047
14	14	Snack Foods	cluster_0	1.780	1.840
2	2	Breads	cluster_1	-0.751	-0.713
3	3	Breakfast	cluster_1	-1.127	-1.089
8	8	Hard Drinks	cluster_1	-0.850	-0.825
12	12	Others	cluster_1	-0.970	-0.980
13	13	Seafood	cluster_1	-1.250	-1.186
16	16	Starchy Foods	cluster_1	-1.026	-0.949
1	1	Baking Goods	cluster_2	0.308	0.121
4	4	Canned	cluster_2	0.310	0.331
5	5	Dairy	cluster_2	0.398	0.422
9	9	Health and ...	cluster_2	-0.034	-0.137
11	11	Meat	cluster_2	-0.287	-0.286
15	15	Soft Drinks	cluster_2	-0.234	-0.315

Gambar4. Hasil Pengujian menggunakan *Rapidminer*

d. Hasil pengukuran performa



PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.192
Avg. within centroid distance_cluster_0: -0.451
Avg. within centroid distance_cluster_1: -0.052
Avg. within centroid distance_cluster_2: -0.159
Davies Bouldin: -0.497
```

Gambar 5. Tampilan hasil ukur di *Rapidminer*

5. Kesimpulan

Dengan adanya pengelompokan data ini, dapat memahami produk yang paling laku, laku, dan kurang laku. Sehingga tidak akan ada lagi penumpukan barang di gudang. *Output* yang dihasilkan dari penelitian ini yaitu, barang yang paling laris menggunakan algoritma *k-meansMicrosoft Excel* paling laku 4 kategori produk, barang yang laku terdapat 6 kategori produk, dan barang kurang laku terdapat 6 kategori produk.Seperti ditunjukkan dalam tabel 9 dibawah ini:

Tabel 9. Klasifikasi akhir kategori produk paling laku (C_0), laku (C_2), dan kurang laku (C_1)

No.	Tipe Barang	Qty	Sales	Tipe	Ket
6	Frozen Foods	856	1,825,735	C_0	PALING LAKU
7	Fruits and Vegetables	1232	2,820,060	C_0	
10	Household	910	2,055,494	C_0	
14	Snack Foods	1200	2,732,786	C_0	

2	Breads	251	553,237	C_1	KURANG LAKU
3	Breakfast	110	232,299	C_1	
8	Hard Drinks	214	457,793	C_1	
12	Others	169	325,518	C_1	
13	Seafood	64	148,868	C_1	
16	Starchy Foods	148	351,401	C_1	
1	Baking Goods	648	1,265,525	C_2	LAKU
4	Canned	649	1,444,151	C_2	
5	Dairy	682	1,522,594	C_2	
9	Health and Hygiene	520	1,045,200	C_2	
11	Meat	425	917,566	C_2	
15	Soft Drinks	445	892,898	C_2	
		8,523	**18,591,125**		

Dalam evaluasi *cluster* yang dilakukan didapatkan nilai *Davies Bouldin Index (DBI)*,yaitu 0,497, sehingga bisa dikategorikan *cluster* yang didapatkan sudah cukup baik karena kemiripan antar anggota *cluster* berada pada >=0.

Daftar Pustaka

Bungin H.M. Burhan Bungin, 2017, *Metodologi Penelitian Kuantitatif, Komunikasi, Ekonomi, dan Kebijakan Publik Serta Ilmu-Ilmu Sosial Lainnya,* Fajar Interpratama Mandiri, Jakarta.

Jollyta Denny Jollyta, et.all, 2021, *Teknik Evaluasi Cluster, Solusi Menggunakan Python dan Rapidminer*, Deepublish Publisher, Yogyakarta.

Novia Emha Ainun Novia, Woro Isti Rahayu, dan Cahyo Prianto, 2020, *Sistem Perbandingan Algoritma K-Means dan Naïve Bayes untuk Mempredriksi Prioritas Pembayaran Tagihan Rumah Sakit Berdasarkan Tingkat Kepentingan.,* Kreatif Industri Nusantara, Bandung.

Prasetyowati Erwin Prasetyowati, 2017, *Data Mining, Pengelompokan Data untuk Informasi dan Evaluasi,* Duta Media Publishing, Pamekasan.

Siregar Amril Mutoi Siregar, Adam Pusphabuana, 2020, *Data Mining, Pengolahan Data Menjadi Informasi dengan Rapid Miner,* CV Kekata Group, Jakarta.

Wahyudi Mochamad Wahyudi, Masitha, Risna Saragih, dan Solikhun, 2020, *Data Mining: Penerapan Algoritma K-Means Clustering dan K-Medoids Clustering*, Yayasan Kita Menulis, Jakarta.

Wanto Anjar Wanto, Mohammad Noor Hasan Siregar, dan Agus Perdana Windarto, 2020, *Data Mining: Algoritma dan Implementasi,* Yayasan Kita Menulis, Jakarta.