IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING UNTUK PENGELOMPOKAN MINAT KONSUMEN PADA PRODUK ONLINE SHOP

Muhammad Ali Khofi Lutfi[1], *Agung Nilogiri*[2]
m1198ali@gmail.com,
agungnilogiri@unmuhjember.ac.id

The development of internet technology is currently very rapid in various fields, including in the business world. this can be seen with the emergence of various businesses in the field of sales based online or commonly called online shops. The bigger an Online shop, the more transactions that are carried out, and can attract data that is too large.In sales transaction activities, consumer interest in the sale of a product can be measured by the number of sales transactions carried out. Later this information can be used as a determination of marketing strategies. Then it takes a data mining technique to get various information that is useful for the company. K-Means is one method of non-hierarchical clustering data that partition data into clusters so that data that has the same characteristics are grouped into one and the same cluster of data that has different characteristics grouped into other groups. The results of this study indicate that the optimal number of Clusters is 3 Clusters with DBI value 0.469 and from 2708 sales data there are 50 products which are members of Cluster 1, 355 products from Cluster 2 member and 2303 products from Cluster 3 members. It is expected that this research can be useful for the company and as a reference for future research.

Keyword : Consumer Interest, K-Means Clustering, Online Shop

Perkembangan teknologi internet saat ini begitu pesat dalam berbagai bidang, tak terkecuali dalam dunia bisnis. hal ini dapat dilihat dengan munculnya berbagai usaha dibidang penjualan berbasis online atau biasa disebut Online shop. Semakin besar sebuah Online shop maka semakin banyak transaksi yang dilakukan, serta dapat menarik data yang begitu besar pula. Dalam kegiatan transaksi penjualan, minat konsumen terhadap penjualan suatu produk dapat diukur dari banyaknya jumlah transaksi penjualan yang dilakuan. Nantinya informasi tersebut dapat digunakan sebagai penentuan strategi pemasaran. Maka dibutuhkan sebuah Teknik data mining untuk mendapatkan berbagai informasi yang bermanfaat bagi perusahaan. K-Means merupakan salah satu metode data clustering non hirarki yang mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok lain. Hasil dari penelitian ini menunjukkan bahwa jumlah Cluster paling optimal yaitu 3 Cluster dengan nilai DBI 0.469 dari 2708 data penjualan terdapat 50 produk merupakan anggota Cluster 1, 355 produk anggota Cluster 2 dan 2303 produk anggota Cluster 3. Diharapkan penelitian ini dapat bermanfaat untuk perusahaan dan sebagai acuan untuk penelitian selanjutnya.

Kata Kunci: Minat Konsumen, K-Means Clustering, Online Shop

1. PENDAHULUAN

Perkembangan teknologi internet saat ini begitu pesat dalam berbagai bidang tak terkecuali dalam dunia bisnis, hal ini dapat dilihat dengan munculnya berbagai usaha dibidang penjualan berbasis online atau biasa disebut *electronic commerce*Lembaga atau perusahaan yang mengaplikasikan *E-Commerce* dalam kegiatan pemasaran atau perdaganganya biasa dikenal dengan nama *Online shop*.

Semakin besar sebuah *Online shop* maka semakin banyak transaksi yang dilakukan, serta dapat menarik data yang begitu besar pula. Salah satu informasi yang dapat diproleh yaitu untuk mengetahui minat konsumen pada penjualan produk, dimana minat konsumen terhadap penjualan suatu produk dapat diukur dari banyaknya jumlah transaksi penjualan yang dilakuan. Dalam sebuah transaksi penjualan minat konsumen dapat dibedakan menjadi beberapa kelompok berdasarkan tingkatanya. Maka teknik pengolahan data sangat diperlukan untuk menggali informasi dari data tersebut. Dalam hal ini maka digunakanlah teknik *Data Minig.*

Salah satu metode dalam *data mining* yaitu *Clustering* atau pengelompokan. Dari beberapa teknik *Clustering* yang paling sederhana dan umum dikenal adalah algoritma *K-Means*. *K-Means* merupakan salah satu metode data *Clustering non hirarki* yang mempartisi data ke dalam *Cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *Cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok lain (Agusta, 2007).

2. TINJAUAN PUSTAKA

Online Shop berasal dari dua kata yaitu *Online* dan *Shop*, *Online* yang memiliki arti dalam jaringan atau disingkat daring adalah keadaan disaat seseorang terhubung ke dalam suatu jaringan atau sistem yang lebih besar. sedangkan *Shop* yang memiliki arti toko. Jadi *Online shop* adalah sebuah sarana atau fasilitas yang dibuat sebuah perusahaan untuk memasarkan

produk atau jasa melalui internet baik itu berupa website maupun aplikasi.

Data mining merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Moertini, 2002). Kata *mining* sendiri berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar. Karena itu Data *Mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistika dan *database*. Dengan arti lain *data mining* adalah proses untuk penggalian pola-pola dari data. Data *mining* menjadi alat yang semakin penting untuk mengubah data tersebut menjadi informasi.

Data *mining* adalah analisis otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecendrungan yang penting yang biasanya tidak disadari keberadaannya (Moertini, 2002). Hal-hal penting yang terkait dengan data *mining* adalah (Luthfi & Kusrini, 2009).

1. Data *mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar.

Clustering atau pengklusteran adalah metode penganalisaan data, yang sering dimasukkan sebagai salah satu metode *Data Mining*, yang tujuannya adalah untuk mengelompokkan data dengan karakteristik yang sama ke suatu 'wilayah' yang sama dan data dengan karakteristik yang berbeda ke 'wilayah' yang lain.Ada beberapa pendekatan yang digunakan dalam mengembangkan metode *Clustering*. Dua pendekatan utama adalah *Clustering* dengan pendekatan partisi dan *Clustering* dengan pendekatan hirarki. *Clustering* dengan pendekatan partisi atau sering disebut dengan *partition-based Clustering* mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam *Cluster-Cluster* yang ada.

K-Means (MacQueen, 1967) adalah salah satu dari algoritma *unsupervised learning* yang paling sederhana untuk menyelesaikan masalah *Clustering* yang telah dikenal. Prosedur ini mengikuti cara sederhana dan mudah untuk mengklasifikasikan kumpulan data tertentu melalui jumlah *Cluster* tertentu (menganggap *k Cluster*) yang telah ditetapkan sebelumnya. Adapun tujuan dari data *Clustering* ini adalah untuk meminimalisasikan *objective function* yang diset dalam proses *Clustering*, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu *Cluster* dan memaksimalkan variasi antar *Cluster*. Pada dasarnya penggunaan algoritma dalam melakukan proses *Clustering* tergantung dari data yang ada dan konklusi yang ingin dicapai. Untuk itu digunakan algoritma *K-Means* yang didalamnya memuat aturan sebagai berikut:

1. Jumlah *Cluster* perlu diinputkan.
2. Hanya memiliki atribut bertipe numerik.

Davies-bouldin index merupakan salah satu metode yang digunakan untuk mengukur validitas *cluster* pada suatu metode pengelompokan. Pengukuran dengan *davies-bouldin index* ini memaksimalkan jarak inter-*cluster* antara *cluster ci* dan *cj* dan pada waktu yang sama mencoba untuk meminimalkan jarak antar titik dalam sebuah *cluster*. Jika jarak inter-*cluster* maksimal, berarti kesamaan karakteristik antar-masing-masing *cluster* sedikit sehingga perbedaan antar-*cluster* terlihat lebih jelas. Jika jarak intra-*cluster* minimal berarti masing-masing objek dalam *cluster* tersebut memiliki tingkat kesamaan karakteristik yang tinggi (wani & riyaz 2017).

3. METODE PENELITIAN

K-Means Clustering merupakan salah satu metode data *Clustering* non hirarki yang mempartisi data ke dalam *Cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *Cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok lain (Agusta, 2007).

Algoritma *K-Means Clustering* memiliki langkah-langkah berikut :

1. Menentukan banyaknya/jumlah *Cluster* k
2. Menentukan nilai pusat (*Centroid)*. Dalam menentukan nilai *Centroid* untuk awal iterasi, nilai awal *Centroid* ditentukan secara acak. Sedangkan untuk menentukan nilai *Centroid* yang merupakan tahap dari iterasi selanjutnya maka digunakan rumus sebagai berikut :

$$v_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj}$$

Dimana :

V_{ij} : *Centroid* atau rata-rata *Cluster* ke-i untuk variable ke-j
N_i : Jumlah data anggota *Cluster* ke-i
i,k : Indeks dari *Cluster*
X_{kj} : Nilai data ke-k yang ada di dalam *Cluster* tersebut untuk variable ke-j

3. Menghitung jarak antara titik *Centroid* dengan titik tiap objek. Untuk menghitung jarak tersebut dapat menggunakan *Euclidean Distance*, yaitu :

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - y_{jk})^2}$$

Dimana :

d_{ij} : Jarak objek antara objek i dan j
p : Dimensi data
x_{ik} : Koordinat dari objek i pada dimensi k
x_{jk} : Koordinat dari objek j pada dimensi k

4. Pengelompokan objek. Untuk menentukan anggota *Cluster* adalah dengan memperhitungkan jarak minimum objek. Nilai yang diperoleh dalam keanggotaan data pada distance matriks adalah 0 atau 1, dimana nilai 1 untuk data yang dialokasikan ke *Cluster* dan nilai 0 untuk data yang dialokasikan ke *Cluster* yang lain.

5. Kembali ke tahap 2. Lakukan perulangan hingga nilai *Centroid* yang dihasilkan tetap dan anggota *Cluster* tidak berpindah ke *Cluster* lain.

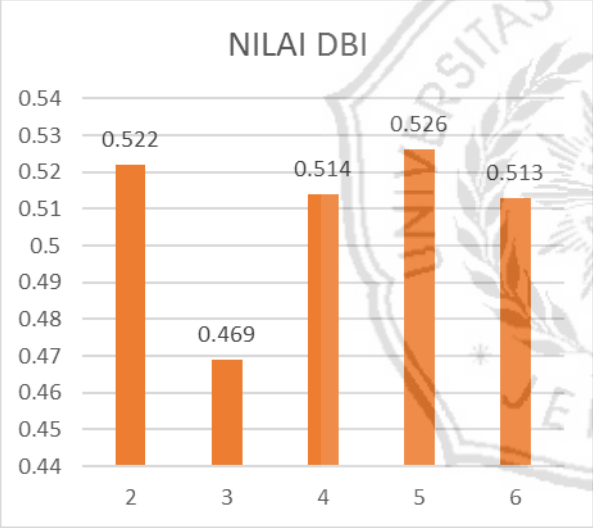
4. HASIL DAN PEMBAHASAN

4.1 Menentukan Jumlah Cluster Paling Optimal

Untuk menentukan jumlah cluster paling optimal penulis menggunakan program bantu *Rapidminer Studio* untuk mendapatkan hasil yang lebih cepat dan akurat. Pengujian dilakukan menggunakan beberapa cluster yang berbeda untuk mendapatkan jumlah cluster paling optimal untuk karakteristik data penjualan *Online Shop* yang akan di implementasikan pada proses selanjutnya.

Data yang diuji pada penelitian ini sebanyak 2708 data. Dimana data tersebut merupakan data penjualan produk pada satu periode penjualan yaitu bulan September 2011.

Dari hasil pengujian performa cluster menggunakan *Davies Bouldin Index* akan dibandingkan dan dicari nilai paling kecil atau nilai yang mendekati 0, Dimana semakin kecil nilai DBI maka semakin optimal *Cluster* yang dibutuhkan.



Gambar 3.1 Perbandingan Nilai DBI

Dari beberapa kali pengujian menggunakan 5 *Cluster* yang berbeda menunjukkan bahwa nilai DBI terkecil terdapat pada pengujian menggunakan 3 jumlah *Cluster* yaitu 0.469 yang berarti implementasi algoritma *K-Means Clustering* untuk pengelompokan minat konsumen pada data penjualan *Online Shop* ini paling optimal menggunakan 3 *Cluster*.

3.1 Implementasi K-Means Clustering

1. Input Data

   Pada penelitian ini akan diinputkan sebanyak 2708 data berupa jenis produk yang memiliki 3 atribut yaitu jumlah transaksi, total penjualan dan rata-ratadiambil dari jumlah transaksi dan total penjualan masing-masing produk selama periode tertentu. Periode yang digunukan pada penelitian ini adalah periode transaksi selama 1 bulan yaitu pada bulan Septermber 2011.

Tabel 3.1 Data yang Diproses

NO	PRODUK	TRANSAKSI	PENJUALAN	RATA-RATA
1	4 PURPLE FLOCK..	3	9	3.00
2	50S CHRISTMAS ..	22	331	15.05
3	DOLLY GIRL BEA..	30	616	20.53
4	I LOVE LONDON ..	26	137	5.27
5	OVAL WALL MIRR..	3	4	1.33
6	RED SPOT GIFT ..	26	376	14.46
7	SPACEBOY BABY ..	20	49	2.45
8	TRELLIS COAT R..	22	121	5.50
9	10 COLOUR SPACE..	28	577	20.61
10	12 COLOURED PAR..	13	151	11.62
...	...	...	...	...
2708	ZINC WIRE SWEET..	1	2	2.00

2. Menentukan Jumlah Cluster dan Nilai Centroid Awal

   Tahap selanjutnya yaitu menentukan jumlah *Cluster* atau kelompok dan nilai *Centroid* awal. Pada tahap sebelumnya diketahui jumlah *Cluster* paling optimal untuk pengelompokan minat konsumen yaitu 3 *Cluster*. Dan untuk nilai awal masing-masing *Centroid* terbentuk secara otomatis yaitu

Tabel 4.2 Nilai Centroid Awal

Centroid (C)	1	2	3
JUMLAH TRANSAKSI	197.10	17.73	1.10
TOTAL PENJUALAN	3,780.90	200.45	1.10
RATA RATA	334.80	9.73	1.10

3. Menentukan Kelompok dan Menghitung Kembali Nilai Centroid

   Pada tahap ini sistem akan menghitung jarak titik pusat *Cluster* terhadap titik tiap objek (data penjualan) selanjutnya dilakukan pengelompokan berdasarkan perbandingan dan dipilih jarak terdekat antara data dengan pusat *Cluster*, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat *Cluster* terdekat dengan cara membandingkan hasil *Cluster* dan diambil yang terkecil.

4. Iterasi Proses K-Means

   Pada tahap ini sistem akan mengulang proses perhitungan menggunakan nilai titik pusat *Cluster* (*Centroid*) terbaru. Proses ini akan berlangsung secara terus menerus hingga syarat berhenti-nya proses *K-Means Clustering* terpenuhi yaitu saat hasil dari pengelompokan dan nilai *Centroid* yang terbentuk tidak berubah dari hasil perhitungan sebelumnya.

Tabel 4.4 Hasil Pengelompokan pada Iterasi Akhir

NO	PRODUK	TRANSAKSI	PENJUALAN	RATA-RATA	C1 (PKG)	C2 (PKG)	C3 (PKG)	KELOMPOK
1	4 PURPLE FLOCK..	3	9	3.00	2164.78	698.82	71.33	C3
2	50S CHRISTMAS ..	22	331	15.05	1842.14	376.09	251.63	C3
3	DOLLY GIRL BEA..	30	616	20.53	1556.99	92.12	536.77	C2
4	I LOVE LONDON ..	26	137	5.27	2036.03	569.74	59.17	C3
5	OVAL WALL MIRR..	3	4	1.33	2169.81	703.85	76.41	C3
6	RED SPOT GIFT ..	26	376	14.46	1797.04	330.94	296.73	C3
7	SPACEBOY BABY ..	20	49	2.45	2124.21	657.96	32.29	C3
8	TRELLIS COAT R..	22	121	5.50	2052.14	585.89	42.68	C3
9	10 COLOUR SPACE..	28	577	20.61	1596.03	130.81	497.74	C2
10	12 COLOURED PAR..	13	151	11.62	2022.38	556.31	71.43	C3
...	...	...	...	...	...	...	...	...
2708	ZINC WIRE SWEET..	1	2	2.00	2171.87	705.96	78.58	C3

Tabel 4.3 Centroid yang Terbentuk pada Iterasi Akhir

Centroid (C)	1	2	3
JUMLAH TRANSAKSI	88.38	49.46	11.31
TOTAL PENJUALAN	2,171.74	706.04	79.72
RATA RATA	42.21	21.13	7.27

5. Hasil Pengelompokan

Setelah didapatkan hasil akhir dari proses iterasi *K-Means Clustering* diperoleh hasil pengelompokan pada masing-masing *Cluster* yaitu:

Tabel 4.8 Hasil Pengelompokan

CLUSTER	JUMLAH PRODUK
Cluster 1	50
Cluster 2	355
Cluster 3	2303

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Dari hasil penelitian implementasi algoritma *K-Means Clustering* untuk pengelompokan minat konsumen pada produk online shop yang telah diuraikan pada bab-bab sebelumnya dapat disimpulkan bahwa:

1. Dari 5 kali pengujian dan dilakukan evaluasi performa menggunakan metode *Davies Bouldin Index*. Jumlah *Cluster* yang paling optimal untuk data penjualan produk *Online Shop* yaitu 3 *Cluster*. Karena menghasilkan nilai DBI paling rendah atau mendekati 0.
2. Penerapan algoritma *K-Means Clustering* pada data penjualan produk pada *Online Shop*, menghasilkan sebuah informasi mengenai data pengelompokan minat konsumen. Dari 2708 produk yang diteliti terdapat 50 produk yang merupakan anggota *Cluster* 1, 355 produk anggota *Cluster* 2 dan 2303 produk anggota *Cluster*

4.2 Saran

Berdasarkan penelitian yang telah dilakukan peneliti dapat memberikan saran untuk penelitian selanjutnya, antara lain:

1. Untuk mendapatkan hasil yang lebih akurat gunakan data yang memiliki atribut yang lebih spesifik sehingga dapat dihitung berdasarkan kategori tertentu.
2. Penelitian ini dapat dikembangkan dengan metode *data mining* lainnya seperti *C-Means Clustering* Untuk mendapatkan hasil yang lebih variatif.

DAFTAR PUSTAKA

Agusta, Y. 2007. *K-means penerapan, permasalahan dan metode terkait*. Jurnal Sistem dan Informatika Vol.3 : 47-60.

Moertini, V. S. 2002. *Data mining sebagai solusi bisnis*. Integral, vol 7 no.1. 87

Larose, D. T. 2005. *Discovering Knowledge In Data: An Introduction To Data Mining*. Jhon Willey & Sons, Inc.

Kusrini & Luthfi, E. T. 2009. *Algoritma Data Mining*. Yogyakarta: Andi.

Ponniah, P. 2001. *Datawarehouse Fundamentals: A Comprehensive Guide For IT Professional*. Jhon Willey & Sons, Inc.

Piatetsky, G. & Shapiro. 2006. *An Introduction Machine Learning, data mining, and Knowledge discovery*, Course in data mining Kdnuggets.

Dubes, R. C. & Jain, A. K. 1988. *Algorithms For Clustering Data*. New Jersey: Prentice Hall.

Fayyad, U. M. 1996. *Advances In Knowledge Discovery and Data Mining*. Camberidge. MA: The MIT Press.

Hammouda, K. & Karray, F. 2003. *A Comparative Study of Data Clustering Techniques*. Canada: University of Waterloo.

Han, J. and Kamber, M. 2006. *Data Mining Concepts and Techniques Second Edition*. Morgan Kauffman, San Francisco.

MacQueen, J.B. 1967. *Some Methods For Classification and Analysis of Multivariate Observations*. 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press. (pp. 1:281-297).

Wani, M. A. & Riyaz, R. 2017. *A Novel Point Density Based Validity Index For Clustering Gene Expression Datasets*. International Journal of Data Mining and Bioinformatics 17(1): 66–84.