Klasterisasi Produk Berdasarkan Data Penjualan Menggunakan Algoritma K-Means Dengan Penentuan Centroid Awal

Risaldi Istighfariyansyah[1], Maftahatul Hakimah[2], Muchamad Kurniawan[3]
Jurusan Teknik Informatika, Institut Teknologi Adhi Tama Surabaya[1, 2, 3]
e-mail: hakimah.mafta@itats.ac.id

ABSTRACT

The pandemic conditions at the beginning of 2020 made business owners, especially culinary businesses, implement strategies to make their products sell. This happened at one of the kafes in Surabaya where within a period of one year in the 2020 period, the sales level of a product offered was less than optimal. One strategy that can be taken and discussed in this research is to map sales products from those that sell best to those that sell less. This mapping can be done using a clustering approach. This research applies a simple and effective method for clustering, namely K-Means. K-means is a type of partition-based clustering that works by randomly determining the centroid of each cluster and then each instance will be grouped into clusters with the closest distance. However, K-Means method has a drawback, namely determining the initial centroid randomly. So, this research applies K-Means with initial centroid determination. Based on the test results, K-Means with initial centroid determination has increased the Davies Bouldin Index (DBI) value of the K-Means Standard by 71.68% and has reduced the Sum of Squared Error (SSE) value of the K-Means Standard by 35.73 %.

Keywords: *k-means, mean, variance, DBI, initial centroid*

ABSTRAK

Kondisi pandemi pada awal Tahun 2020 membuat pada pemilik usaha terutama usaha kuliner memasang strategi untuk membuat produknya laku. Hal tersebut terjadi di salah satu Kafe di Surabaya dimana dalam kurun waktu satu tahun pada periode tahun 2020, tingkat penjualan suatu produk yang ditawarkan kurang maksimal. Salah satu strategi yang bisa diambil dan dibahas pada penelitian ini adalah memetakan produk penjualan mulai yang paling laku sampai dengan yang kurang laku. Pemetaan tersebut bisa dilakukan dengan pendekatan klasterisasi. Penelitian ini menerapkan metode yang sederhana dan efektif dalam klasterisasi yakni K-Means. K-means merupakan salah satu partition-based clustering yang bekerja dengan cara menentukan secara acak centroid dari tiap cluster kemudian tiap *instance* akan dikelompokkan ke dalam cluster dengan jarak terdekat. Namun Metode K-Means memiliki kekurangan yaitu penentuan centroid awal dengan acak. Sehingga, penelitian ini menerapkan K-Means Dengan Penentuan Centroid Awal. Berdasarkan hasil pengujian, K-Means Dengan Penentuan Centroid Awal bisa meningkatkan nilai Davies bouldin index (DBI) dari K-Means Standart sebesar 71,68% dan dapat menurunkan nilai Sum of Squared Error (SSE) dari K-Means Standart sebesar 35,73%.

Kata kunci: k-means, mean, varians, DBI, centroid awal

PENDAHULUAN

Mengunjungi kafe saat ini menjadi gaya hidup sebagian besar masyarakat kota untuk melepas lelah setelah seharian bekerja. Adanya kenyataan ini, bisnis kafe menjadi sebuah ide usaha dengan modal relatif kecil. Kafe sendiri merupakan suatu restoran informal dengan mengutamakan pada penyajian makanan, tempat yang nyaman untuk bersantai. Dalam dunia bisnis, keuntungan menjadi tujuan penting dalam operasional sebuah usaha, sehingga perlu strategi untuk dapat meningkatkan penjualan dan dapat menarik minat konsumen untuk berkunjung pada kafe. Hal ini juga terjadi pada salah satu Kafe yang ada di Surabaya. Tingkat penjualan kafe tersebut mengalami penurunan selama pandemi pada Tahun 2020.

Penelitian ini membantu pihak Kafe mengatasi permasalahan dengan memetakan produk penjualan mulai yang paling laku hingga yang tidak laku. Pemetaan tersebut dilakukan

menggunakan konsep klasterisasi. Salah satu metode yang sangat sederhana dan efektif dalam proses clustering adalah K-Means. Metode ini merupakan salah satu *partition-based clustering* yang bekerja dengan cara menentukan secara acak centroid dari tiap cluster kemudian tiap *instance* akan dikelompokkan ke dalam cluster dengan jarak terdekat[1]. Keunggulan dari metode K-means adalah karena implementasi yang mudah dan sederhana, skalabilitas, kecepatan konvergensi, serta adaptasi terhadap data[2]. Namun, metode K-Means memiliki kekurangan pada penentuan centroid awal[3].

Kelemahan metode k-means dalam penentuan centroid awal secara random kemudian diperbaiki dan dipublikasikan pada penelitian[4]–[8]. Pada penelitian [4], inisialisasi centroid dicari dengan menerapkan metode *single linkage*. Pada penelitian [5], centroid awal ditentukan dari means dari fitur dataset. Sementara itu, penelitian [6] mengusulkan penentuan centroid awal berdasarkan vektor eigen dari suatu matriks baru yang dibentuk dari matriks dataset yang dikali dengan transpose dari matriks tersebut. Perbaikan metode k-means juga diusulkan pada Penelitian [7] dengan memberikan bobot pada fitur dataset. Untuk penentuan centroid awal diusulkan menggunakan means dan varians yang dihitung menggunakan Rumus yang disajikan pada penelitian tersebut. Sedangkan, penentuan centroid awal pada penelitian [8] diperoleh dengan menerapkan pencarian heuristik menggunakan Genetic Algorithm Polygamy. Konsep yang digunakan adalah memasangkan 1 *father* terpilih dengan lebih dari 1 *mother* untuk mendapatkan individu terbaik. Dari beberapa penelitian yang dilakukan sebelumnya, klasterisasi produk pada penelitian ini menggunakan *k*-means dengan penentuan centroid awal menggunakan Persamaan yang diusulkan pada penelitian [7].

TINJAUAN PUSTAKA

K-Means

K-means clustering merupakan metode non-hirarki yang dapat mengelompokan sebuah objek menjadi satu cluster atau lebih. Objek data dengan karakteristik sama dapat dikelompokan dalam satu cluster yang sama dan sebaliknya, data yang memiliki karakteristik berbeda dimasukkan dalam cluster yang lain. Dengan proses tersebut, objek data yang berada dalam satu cluster memiliki tingkat variasi yang sangat kecil[9]. Perpanjangan alami dari masalah K-Means memungkinkan kita untuk memasukkan beberapa informasi lebih lanjut. Ini mungkin mewakili ukuran kepentingan, hitungan frekuensi, atau lainnya informasi. K-Means mencoba untuk menguraikan satu set objek menjadi satu set cluster yang terputus-putus, dengan mempertimbangkan fakta bahwa atribut numerik objek dalam himpunan sering tidak berasal dari distribusi normal identik yang independen.

K-Means Dengan Penentuan Centroid Awal

Penentuan centroid awal pada penelitian ini menggunakan Persamaan yang diusulkan pada Penelitian [7] dengan bobot fitur pada penelitian ini dianggap sama. Algoritma K-means dengan penentuan centroid awal ini diberikan sebagai berikut.

1. Tentukan centroid awal dengan Persamaan berikut :

$$C = \begin{cases} \left\{\bar{x} \pm \frac{2v}{k-1} \times j, j = 1,2, \dots, k/2\right\} \cup \{mean\}, & ketika\ k\ ganjil \\ \left\{\bar{x} \pm \frac{2v}{k} \times j, j = 1,2, \dots, k/2\right\}, & ketika\ k\ genap \end{cases} \quad (1)$$

dengan, C = centroid,
v = Varian
$\bar{x}$ = mean/(rata-rata)
x_i= data ke i

n = banyak data
k = Jumlah cluster

Dalam persamaan tersebut dimana, 2 v/k dan 2v/(k-1) adalah faktor offset. Jika jumlah sampel adalah nol setelah iterasi pertama, faktor offset terlalu besar, dan kemudian dibelah dua untuk memilih kembali cluster pusat.

2. Menghitung jarak setiap objek ke pusat kluster ($d(x,y)$) dalam data set menggunakan :

$$d(x,y) = \sqrt{\sum(x_i - y_i)^2} \quad (2)$$

Cluster ditetapkan untuk setiap objek berdasarkan jarak minimum.
3. Hitung rata-rata untuk setiap cluster, kemudian update centroid dengan nilai rata-rata tersebut.
4. Langkah-langkah (2) dan (3) diulang sampai iterasi berakhir atau hasil pengelompokan tidak lagi berubah.

Evaluasi Metode Cluster

Untuk mengukur kinerja metode clustering dalam penelitian ini menggunakan *Davies Bouldin Index* (DBI) dan *Sum of Squared Error* (SSE). DBI sendiri merupakan salah satu pengujian yang dapat membantu dalam mengukur akurasi cluster. Pengukuran DBI bertujuan untuk memaksimalkan jarak antara satu cluster dengan cluster lain dan meminimalkan jarak antar objek data yang terdapat dalam cluster yang sama. Sedangkan SSE bertujuan untuk mencari nilai selisih yang kecil dari jarak antara satu cluster dengan cluster lain.

Davies Bouldin Index (DBI)

Davies-bouldin index adalah salah satu metode evaluasi internal untuk mengukur sebuah cluster pada suatu metode clustering yang didasarkan pada nilai kohesi dan separasi. Dalam clustering, kohesi merupakan jumlah kedekatan objek data terhadap centroid suatu cluster. Sedangkan separasi diukur berdsarkan jarak antar centroid pada masing-masing cluster.

Sum of square within cluster (SSW) merupakan persamaan yang dapat digunakan untuk mengetahui sebuah matrik kohesi dalam suatu cluster ke-*i* pada Persamaan berikut ini.

$$SSW_i = \frac{1}{m_i}\sum_{j=1}^{m_i} d(x_j, c_i) \quad (3)$$

dengan, mi merupakan jumlah data dalam cluster ke-*i*, c_i adalah centroid cluster ke-*i*, dan $d(x_j, c_i)$ adalah jarak setiap data terhadap centroidnya menggunakan jarak euclidean. *Sum of square between cluster* (SSB) merupakan persamaan yang digunakan untuk mengetahui sebuah separasi di antar cluster menggunakan persamaan 4 berikut ini :

$$SSB_{i,j} = d(c_i, c_j) \quad (4)$$

Dari Persamaan 3 dan 4, akan dihitung rasio (Rij) untuk mengetahui sebuah nilai perbandingan antara cluster ke-*i* dan cluster ke-*j*. Cluster yang baik adalah cluster yang memiliki nilai kohesi sekecil mungkin dan nilai separasi yang sebesar mungkin. Nilai rasio dihitung menggunakan Persamaan 5.

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{ij}} \quad (5)$$

Nilai rasio pada Persamaan 5 digunakan untuk mencari nilai (DBI) dengan Persamaan sebagai berikut[10].

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(R_{ij}) \quad (6)$$

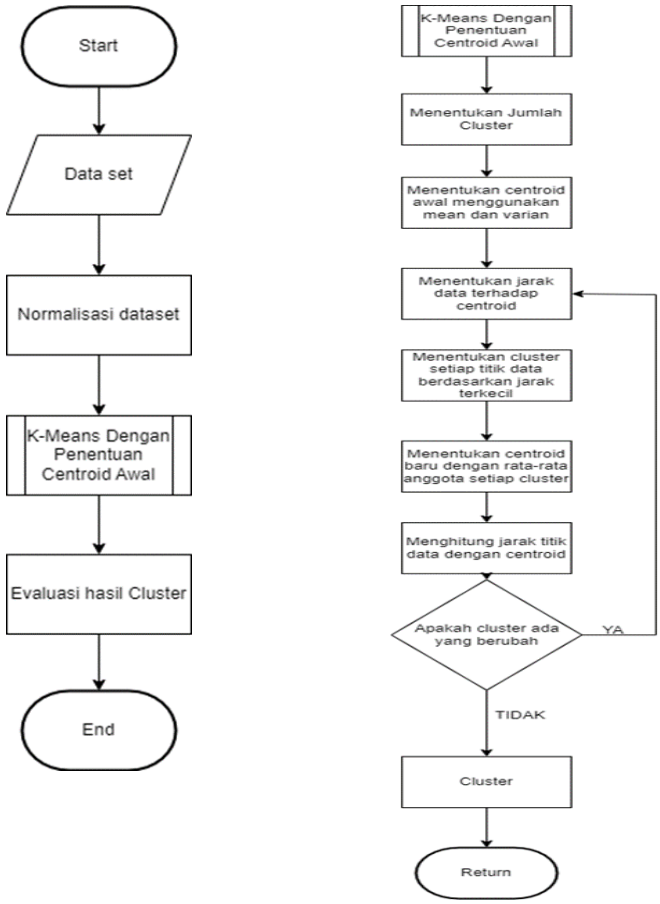
***Sum Of Squared Error* (SSE)**

Sum of Squared Error (SSE) dihasilkan dari total jarak masing-masing obyek data dengan titik pusat clusternya. Semakin kecil nilai SSE, obyek didalam data semakin homogin. Nilai SSE dihitung menggunakan Persamaan 7 berikut ini[11].

$$SSE = \sum_{i=1}^{n} \left(x_{ij} - x_i\right)^2 \quad (7)$$

METODE

Tahapan penelitian untuk mencapai tujuan penelitian diatas dijelaskan pada skema Gambar 1 berikut ini.



Gambar 1 Tahapan Penelitian

Penelitian ini diawali dengan memasukkan dataset penelitian. Dataset berupa harga dan volume penjualan kafe selama setahun. Setelah data siap diolah, langkah selanjutnya adalah normalisasi nilai pada setiap variabel agar seragam. Langkah berikutnya proses klasterisasi dijalankan dengan algoritma yang telah dijelaskan pada bagian Algoritma K-Means dengan penentuan centroid awal. Jumlah cluster pada penelitian ini ditentukan sebanyak 3 cluster. Jika iterasi algoritma k-means dengan penentuan centroid awal telah selesai dan menghasilkan cluster, berikutnya hasil dievaluasi menggunakan Persamaan DBI dan SSE.

HASIL DAN PEMBAHASAN

Data Penelitian

Data penelitian diambil dari salah satu Kafe yang berlokasi di Surabaya. Data set berupa item makanan yang disertai dengan harga dan volume penjualan pada Tahun 2020. Data tersebut disajikan pada Tabel 1 berikut ini.

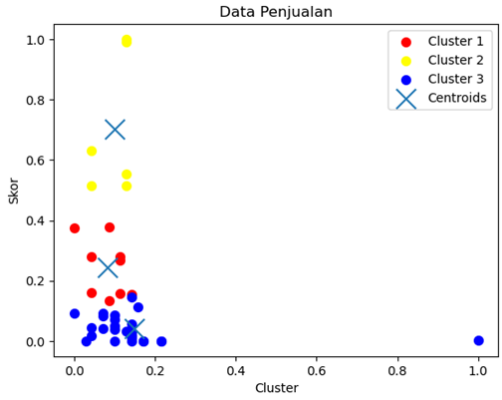
Tabel 1. Data Penelitian

No	Item	Harga Satuan (Rp)	Volume Penjualan			
			Januari	Februari	Maret	April
1	p/w ori	23.000	135	152	148	136
2	p/w banut	29.000	485	500	481	492
3	p/w strawberry	26.000	183	181	169	204
4	p/w straw cheese	23.000	306	380	290	285
5	p/w blue cheese	23.000	250	287	295	270
6	p/w lotus crumble	30.000	9	25	14	11

Dengan dataset pada Tabel 1. Akan dikelompokkan menjadi 3 cluster untuk merepresentasikan item makanan yang termasuk laku, rata-rata dan kurang laku.

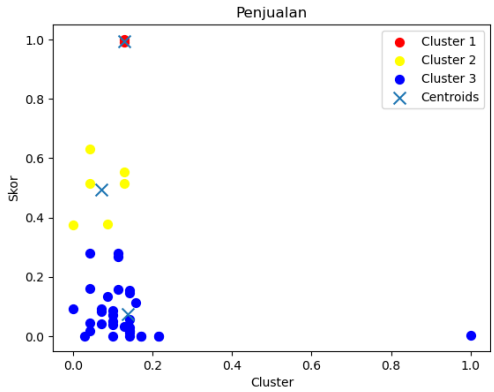
Hasil Klasterisasi menggunakan K-Means Dengan Penentuan Centroid Awal

K-Means dengan penentuan centroid awal untuk hasil uji coba yang di mana untuk menentukan centroid menggunakan rumus mean dan varian yang ada di Persamaan 1 untuk menghasilkan centroidnya. Berikut ini adalah visualisasi dataset hasil cluster menggunakan K-Means Dengan Penentuan Centroid Awal dengan menggunakan tiga centroid. Visualisasi data tersebut menggunakan nilai yang telah dinormalisasi hasil gambar dapat di lihat pada gambar 4.2.



Gambar 2. Visualisasi K-Means dengan penentuan centroid awal.

Hasil klasterisasi k-means dengan penentuan centroid awal bisa dibandingkan dengan hasil klasterisasi dataset menggunakan k-means standart yang diberikan pada Gambar 3 berikut.



Gambar 3. Visualisasi Hasil K-Means K-Means standart

Pengujian hasil klasterisasi dua metode tersebut diberikan pada Tabel 2 sebagai berikut.

Tabel 2. Tabel Hasil SSE dan DBI

Metode	Evaluasi	
	SSE	DBI
K-Means Standart	18.883048658518444	0.3873898851781145
K-Means dengan Penentuan Centroid Awal	12.134444736130414	0.6650807657626223

Hasil pengujian berdasarkan SSE dan DBI menunjukkan bahwa centroid awal menggunakan nilai mean dan varian yang diberikan pada Persamaan 1 bisa meningkatkan kinerja klasterisasi dari k-means standart.

KESIMPULAN

Kesimpulan dari penelitian ini anatara lain sebagai berikut : Metode K-Means dengan penentuan centroid awal mengguan untuk melakukan clustering data penjualan. Hal ini dapat dilakukan dengan cara mengkatgorikan penjulan laku, sedang dan kurang laku. Metode tersebut menunjukkan hasil clustering yang baik untuk hasil yang diketahui bahwa K-Means dengan penentuan centroid awal cukup baik dibandingkan dengan K-Means Standart karena dari hasil pengujian Sum of Squared Error (SSE) K-Means dengan penentuan centroid awal lebih kacil hasil SSEnya dengan penurunan SSE 35,73% dan dari Davies bouldin index (DBI) K-Means dengan penentuan centroid awal lebih lebih besar nilai DBInya dengan peningkatan DBI 71,68%.

DAFTAR PUSTAKA

[1] Q. Han and E. Al., “Vector partitioning quantization utilizing K-means clustering for physical layer secret key generation.,” *Inf. Sci. (Ny).*, vol. 512, pp. 137–160, 2020.

[2] S. N. Gama, I. Cholissodin, and M. T. Furqon, “Clustering Portal Jurnal International untuk Rekomensari Publikasi Berdasarkan Kualitas Cluster Menggunakan Kernel K-Means,” *Progr. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 5, no. 1, 2015.

[3] E. Ongko, “Perbaikan Performance K-Means Melalui Perbaikan Penentuan Centroid,” *J. Ilm. CORE IT*, vol. 9, no. 2, 2020.

[4] A. Aprilia, W. M. Rahmawati, and M. Hakimah, “Penentuan Kategori Status Gizi Balita Menggunakan Penggabungan Metode Klasterisasi Agglomerative Dan K-Means,” *Semin. Nas. Sains dan Teknol. Terap. VII - Inst. Teknol. Adhi Tama Surabaya*, pp. 595–600, 2019.

[5] M. A. Lakshmi, G. V. Daniel, and D. S. Rao, "Initial centroids for k-means using nearest neighbors and feature means," 2019.
[6] S. Manochandar, M. Punniyamoorthy, and R. K. Jeyachitra, "Development of new seed with modified validity measures for k-means clustering," *Comput. Ind. Eng.*, vol. 141, 2020.
[7] Y. Yu, S. A. Velastin, and F. Yin, "Automatic grading of apples based on multi-features and weighted K-means clustering algorithm," *Inf. Process. Agric.*, vol. 7, no. 4, pp. 556–565, 2020, doi: 10.1016/j.inpa.2019.11.003.
[8] R. R. Muhima and E. Al., "An improved clustering based on K-means for hotspots data," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 31, no. 2, pp. 1109–1117, 2023.
[9] Y. Agusta, "K-means–penerapan, permasalahan dan metode terkait," *J. Sist. dan Inform.*, vol. 3, no. 1, pp. 47–60, 2007.
[10] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015, doi: 10.1007/s40745-015-0040-1.
[11] E. Umargono, J. E. Suseno, and V. G. S. K., "K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median," vol. 474, no. Isstec 2019, pp. 234–240, 2020, doi: 10.5220/0009908402340240.