ISSN: 1829-667X / E-ISSN: 2460-9021
DOI: 10.31315/telematika.v22i2.14468

Implementation of Forgy Initialization and K-Means++ Algorithms in the K-Means Clustering Method for Sales Data Analysis of Dazzle Store

Penerapan Algoritma Forgy Initialization dan K-Means++ Dalam Metode K-Means Clustering Untuk Analisis Data Penjualan Toko Aksesoris Dazzle

Muhamad Hilmi Abdillah[1], Frans Richard Kodong[2]
[1,2] Informatika, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

[1]123190069@student.upnyk.ac.id, [2]frans.richard@upnyk.ac.id

Informasi Artikel
Received: January 2025
Revised: February 2025
Accepted: June 2025
Published: June 2025

Keywords:Forgy Initialization; Kmeans++; silhouette coefficient; Kmeans Clustering
Kata kunci: Forgy Initialization; Kmeans++; silhouette coefficient; Kmeans Clustering

Abstract

Objective: To determine the results of K-Means Clustering calculations by applying K-Means++ and Forgy initialization methods in analyzing sales data at Dazzle accessory store, as well as to identify the optimal number of clusters using the silhouette coefficient.
Method: This study implements the Forgy initialization and K-Means++ algorithms in the K-Means Clustering method, along with an evaluation of the optimal number of clusters using the silhouette coefficient method.
Results: The application of Forgy initialization and K-Means++ successfully improved clustering outcomes more optimally compared to the pure initialization method. The highest silhouette coefficient evaluation score was 0.9232095222373023 for K-Means++ and 0.8822890619277 for Forgy initialization. This result is clearly better than the pure initialization method, which only achieved a score of 0.8816344025002508.
State of the Art: This study builds upon previous research. The innovation lies in the implementation of a combination of K-Means Clustering with Forgy initialization and K-Means++ initialization methods.

Abstrak

Tujuan: Mengetahui hasil perhitungan metode K-Means Clustering dengan menerapkan K-Means++ dan Forgy initialization dalam menganalisis data penjualan di toko aksesoris Dazzle, serta mengetahui jumlah cluster optimal dengan *silhouette coefficient.*
Metode: Penelitian ini mengimplementasikan algoritma *forgy initialization* dan *Kmeans++* dalam metode *Kmeans Clustering,* serta evaluasi jumlah cluster optimal menggunakan metode *silhouette coefficient.*
Hasil: Penerapan *forgy initialization* dan *Kmeans++* mampu meningkatkan hasil akhir *clustering* dengan lebih

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI: 10.31315/telematika.v22i2.14468

optimal daripada metode inisiasi murni. Nilai evaluasi *silhouette coefficient* tertinggi sebesar 0.9232095222373023 untuk *Kmeans*++ dan 0.8822890619277 Untuk *forgy initialization.* Hal tersebut jelas lebih baik daripada inisiasi menggunakan metode murni yang hanya mendapat nilai sebesar 0.8816344025002508.
State of The Art: Penelitian ini merupakan pengembangan dari penelitian terdahulu. Inovasi penelitian ini terletak pada penerapan kombinasi *Kmeans Clustering* dengan metode inisiasi *forgy initialization* dan *Kmeans*++.

1. Pendahuluan

Setiap perusahaan disektor perdagangan memiliki keinginan untuk optimal dalam mengembangkan bisnisnya agar dapat tetap bersaing dengan efektif di tengah persaingan bisnis yang sangat ketat[1]. Semakin ketatnya persaingan dalam dunia bisnis mendorong perusahaan-perusahaan untuk intensif dalam meningkatkan performa karyawan mereka. Hal ini dipicu oleh peran krusial teknologi dalam mengoptimalkan efisiensi waktu kerja dan akurasi pengolahan data. Oleh karena itu, banyak perusahaan yang beralih dari sistem manual ke sistem digital guna mengatasi lonjakan data perusahaan yang terus meningkat setiap tahun. Tantangan utama terletak pada pengelolaan bank data, terutama dalam bagian penjualan, yang jika tidak ditangani dengan baik dapat berdampak fatal bagi kelangsungan perusahaan[2].

Contoh kasus yaitu pada toko aksesoris Dazzle, dimana jumlah penjualan yang naik turun mengakibatkan stok barang menjadi kurang stabil dan berpengaruh terhadap konsumen. Stok barang yang tidak dikelola dengan baik juga bisa berdampak pada toko aksesoris Dazzle seperti stok habis ketika permintaan konsumen tinggi yang membuat transaksi barang harus dibatalkan dan akhirnya konsumen membeli produk di tempat yang lain. Kesalahan dalam memprediksi stok barang juga dapat membuat barang tidak habis terjual menjadi menumpuk dan mempersempit ruang di toko. Data penjualan pada toko Dazzle yang setiap bulannya fluktuatif sehingga perlu adanya pengelompokan (Clustering). Salah satu metode yang digunakan untuk clustering adalah K–Means Clustering yang mana memiliki keunggulan dalam kemampuan untuk mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien. Akan tetapi, metode ini juga memiliki kelemahan yang disebabkan oleh penentuan pusat awal cluster. Hasil dari proses clustering dengan K-Means sangat bergantung pada inisiasi nilai pusat awal cluster yang diberikan[3].

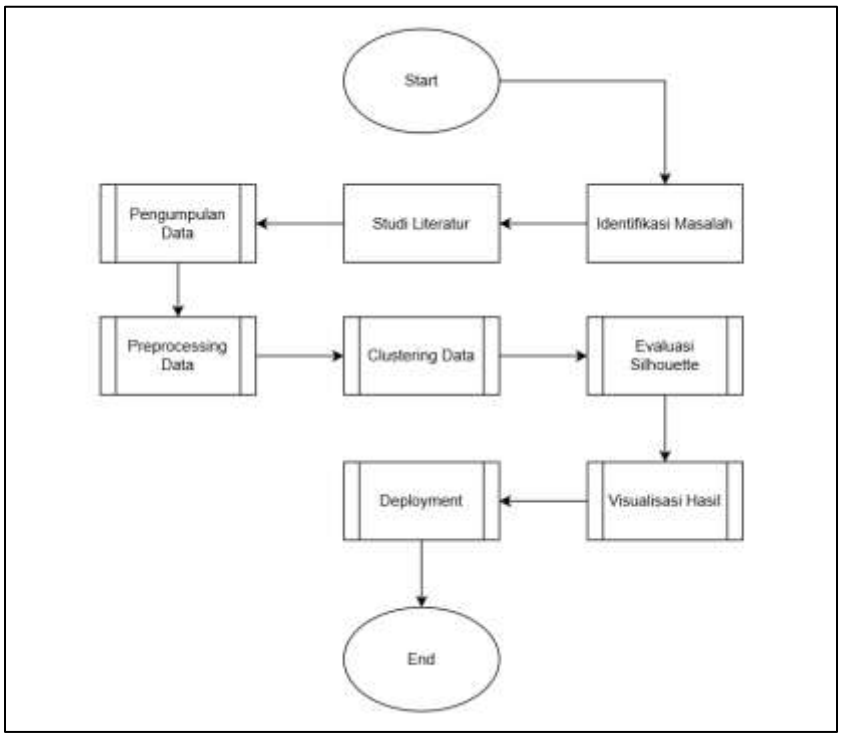
Untuk mengatasi hal tersebut, maka pada penelitian ini akan ditambahkan algoritma Forgy Initialization dan K-Means++. Keduanya merupakan sebuah teknik inisiasi centroid awal yang dapat diterapkan pada metode K-Means Clustering. Kemudian untuk menganalisis jumlah cluster yang optimal, pada penelitian ini akan digunakan metode Silhouette coefficient. Metode Silhouette coefficient digunakan untuk melihat kualitas dan kekuatan cluster, seberapa baik atau buruknya suatu obyek ditempatkan dalam suatu cluster[4].

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI: 10.31315/telematika.v22i2.14468

Penelitian ini bertujuan untuk membagi data penjualan menjadi beberapa klaster dengan menambahkan algoritma Forgy Initialization dan K-Means++ untuk menginisiasi nilai awal pada klaster, serta menggunakan metode Silhouette coefficient untuk menentukan jumlah klaster optimal yang digunakan.

2. Metode/Perancangan

Penelitian ini akan membuat sebuah sistem untuk melakukan klasterisasi data penjualan dengan menambahkan algoritma *Forgy Initialization* dan *K-Means++* untuk menginisiasi nilai awal pada klaster, serta menggunakan metode *Silhouette coefficient* untuk menentukan jumlah klaster optimal yang digunakan.



Gambar 1. Perancangan Proses

2.1. Dataset

Dataset pada penelitian ini merupakan sebuah data primer yang diperoleh secara langsung dari toko aksesoris Dazzle. Data diambil mulai dari tanggal 1 Februari 2024 hingga 8 Februari 2024. Total keseluruhan data yang diperoleh yaitu sejumlah 12.301.

2.2. *Data Preprocessing*

Pembersihan data merupakan proses eliminasi *noise*, perbaikan data yang tidak konsisten, dan pengisian nilai yang hilang, seperti data yang tidak lengkap, duplikasi, kesalahan pengetikan, kehilangan huruf, atau kelebihan huruf, serta masalah lainnya.berlanjut[5]. Terdapat beberapa tahapan dalam proses ini yaitu sebagai berikut:

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI: 10.31315/telematika.v22i2.14468

1. Cleaning Data

Pada proses ini data dibersihkan agar dapat digunakan sesuai dengan kebutuhan penelitian, diantaranya yaitu menghapus baris yang memiliki NaN pada tabel kuantitas dan harga, dan mengubah tabel tersebut menjadi tipe integer.

2. *Transform Data*

Proses awal transformasi data dilakukan dengan membentuk kolom baru bernama total_order. Kolom total_order tersebut merepresentasikan total pesanan untuk setiap produk dalam transaksi. dibentuk berdasarkan berapa banyak kemunculan kode_barang yang sama didalam no_transaksi yang berbeda. Data yang sudah melalui seluruh tahapan preprocessing selanjutnya dilakukan standarisasi data dengan rumus sebagai berikut:

$$X_{standar} = \frac{X - \mu}{\sigma}$$

Keterangan:

X = nilai asli dari fitur

μ = rata – rata dari fitur

σ = standar deviasi dari fitur

2.3. *Clustering Data*

Proses clustering data diawali dengan penentuan jumlah cluster, kemudian melakukan inisialisasi centroid atau pusat awal cluster. Setelah itu, hitung jarak setiap data ke masing-masing centroid. Tentukan data mana yang paling dekat dengan setiap centroid berdasarkan perhitungan jarak tersebut. Kelompokkan data dengan data lain yang berada di cluster yang sama. Selanjutnya, hitung kembali posisi centroid menggunakan rata-rata dari data dalam cluster yang sama. Ulangi proses ini sampai tidak ada perubahan dalam pembagian cluster.

Untuk proses clustering data terbagi menjadi 3 (tiga) kategori yang terdiri sebagai berikut:

a) K-Means Clustering Murni

Proses clustering murni dijalankan secara biasa sesuai dengan yang sudah dipaparkan paragraf sebelumnya tanpa penambahan apapun. *Clustering* murni dilakukan karena digunakan sebagai dasar untuk membandingkan proses clustering dengan penambahan metode lainnya.

b) *K-Means Clustering* dengan *Forgy Initialization*

K-Means Clustering dengan metode *Forgy Initialization* dijalankan secara berbeda dengan metode murni. Perbedaan tersebut dapat dilihat pada bagian inisialisasi centroid. Untuk metode murni inisialisasi *centroid* dilakukan secara random tanpa batasan apapun, sedangkan penambahan metode *Forgy Initialization* ini berguna untuk membatasi nilai

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI: 10.31315/telematika.v22i2.14468

random tersebut agar titik centroid yang didapatkan terbatas hanya yang terdapat pada titik data yang ada saja.

c) *K-Means Clustering* dengan *K-Means++*

Proses K-Means Clustering dengan penambahan K-Means++ memiliki tingkat keakuratan yang lebih tinggi lagi. Hal tersebut dapat terjadi karena penambahan metode K-Means++ menjadikan proses clustering lebih kompleks, dimana pada proses inisialisasi centroid dilakukan perhitungan berdasarkan probabilitas dari jarak yang terdekat.

2.4. Evaluasi *Silhouette Coefficient*

Silhouette coefficient berguna untuk mengoptimalkan sebuah cluster, tujuannya adalah agar pengelompokkan tersebut dianggap baik dan optimal. *Silhouette coefficient* diperoleh dengan menghitung jarak antara objek – objek dalam cluster menggunakan rumus jarak, yakni *Euclidean distance*[6].

Untuk melakukan perhitungan *Silhouette coefficient* maka perhitungan yang akan digunakan adalah menggunakan persamaan sebagai berikut:

$$S(i) = \frac{b(i) - a(i)}{\max(bi - ai)}$$

Keterangan:

a_i = rata – rata jarak objek i dengan seluruh objek dalam satu klaster

b_i = rata – rata objek i dengan objek pada klaster berbeda

Rentang nilai yang dihasilkan oleh perhitungan *Silhouette coefficient* adalah dari -1 hingga 1. Nilai rata-rata *Silhouette coefficient* untuk seluruh data dalam sebuah *cluster* mencerminkan akurasi dari pengelompokan data tersebut. Semakin mendekati nilai 1, struktur pengelompokan data dianggap lebih tepat, sedangkan jika mendekati -1, struktur pengelompokan cenderung tumpang tindih atau *overlapping*[7].

3. Hasil dan Pembahasan

Pada bagian hasil dan pembahasan ini, akan dijabarkan mengenai hasil implementasi dari rancangan yang sudah dibuat.

3.1. Hasil

Hasil *clustering* dari penerapan algoritma *Forgy Initialization* dan *Kmeans++* dengan penentuan 3 (tiga) *cluster* adalah sebagai berikut.

Tabel 1. Hasil Inisiasi Murni

Acuan	Cluster 1	Cluster 2	Cluster 3
Total Produk	205	71	1.897

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI: 10.31315/telematika.v22i2.14468

Rata – rata Total Order	13,1561	9,4225	2,4307
Rata – rata Kuantitas	56,0439	376,409	395,965

Tabel 2. Hasil Inisiasi *Forgy Initialization*

Acuan	Cluster 1	Cluster 2	Cluster 3
Total Produk	249	1.851	86
Rata – rata Total Order	11,7229	2,3052	22,6512
Rata – rata Kuantitas	49,6707	2,8455	395,965

Tabel 3. Hasil Inisiasi *Kmeans++*

Acuan	Cluster 1	Cluster 2	Cluster 3
Total Produk	84	1.875	227
Rata – rata Total Order	21,7381	2,3701	12,6167
Rata – rata Kuantitas	403,928	3,0944	52,6696

Pada **tabel 2** dan **tabel 3** merupakan hasil *K-Means Clustering* beserta dengan penerapan metode *Forgy Initialization* dan *Kmeans++* untuk inisialisasi *centroid* awal. Kemudian setelah hasil dari *clustering* didapatkan selanjutnya dilakukan proses evaluasi dengan menggunakan metode *silhouette coefficient*. Berikut merupakan hasil evaluasi dengan menggunakan perhitungan *silhouette coefficient*.

Tabel 4. Hasil Evaluasi *Silhouette Coefficient*

No	Metode Inisiasi	Jumlah Cluster	*Silhouette Coefficient*
1	*K-Means* Murni	2	0.8816344025002508
2	Forgy Initialization	2	0.8822890619277
3	*K-Means++*	2	0.9232095222373023

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI: 10.31315/telematika.v22i2.14468

4	*K-Means* Murni	3	0.7302506636732323
5	Forgy Initialization	3	0.7369882017715667
6	*K-Means*++	3	0.8569351232663325
7	*K-Means* Murni	4	0.5738820203878838
8	Forgy Initialization	4	0.5823959855174015
9	*K-Means*++	4	0.8675671468701301

Tabel 4 menunjukkan hasil evaluasi *clustering* dengan menggunakan perhitungan metode *silhouette coefficient.* Metode *silhouette coefficient* memiliki nilai berkisar antara -1 hingga 1. Nilai mendekati 1 (positif) menunjukkan bahwa titik data berada jauh dari klaster tetangga, yang mengindikasikan hasil pengelompokan yang baik. Kemudian nilai nol menunjukkan adanya tumpang tindih antara klaster atau titik data yang secara sama mendekati beberapa kluster. Sedangkan nilai mendekati -1 (negatif) memiliki arti bahwa titik data mungkin telah salah dikelompokkan.

3.2. Pembahasan

Tanpa metode inisialisasi tertentu, centroid awal dipilih secara acak di antara titik-titik data. Ini dapat menghasilkan hasil clustering yang berbeda-beda setiap kali algoritma dijalankan karena tergantung pada titik awal yang dipilih, oleh karena itu pada penelitian ini menggunakan algoritma *Forgy Initialization* dan *K-Means*++ untuk mengoptimasi hasil dari *Kmeans Clustering.* Untuk hasil evaluasi *clustering* yang paling optimal didapatkan oleh metode inisiasi *Kmeans*++ dengan penentuan 2 (dua) *cluster* sebesar 0.9232095222373023. sedangkan nilai terendah didapatkan metode inisiasi *Forgy Initialization* dengan 4 (empat) *cluster* sebesar 0.5823959855174015.

4. Kesimpulan dan Saran

Penelitian ini berhasil mendapatkan metode inisialisasi centroid yang paling optimal. Berdasarkan evaluasi menggunakan *silhouette coefficient* hasil yang paling optimal yaitu penerapan metode *Kmeans*++. Hal itu dapat diketahui berdasarkan hasil yang didapatkan, metode *Kmeans*++ selalu mendapatkan hasil evaluasi tertinggi tanpa terpengaruh jumlah *cluster* yang digunakan. *Kmeans*++ dapat menjadi solusi ketika hasil dari *clustering* menggunakan metode *Kmeans* murni tidak optimal dikarenakan inisialisasi centroid awal. Saran untuk penelitian ini yaitu perlu ditambahkan parameter untuk perhitungan sehingga hasil yang didapatkan menjadi lebih valid. Selain itu, perlu dilakukan penelitian dengan metode yang berbeda sehingga penelitian ini dapat dijadikan pembanding dan dapat diketahui metode yang paling optimal untuk digunakan selanjutnya.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI: 10.31315/telematika.v22i2.14468

Daftar Pustaka

[1] Y. Dharma Putra, M. Sudarma, and I. B. A. Swamardika, "Clustering History Data Penjualan Menggunakan Algoritma K-Means," *Maj. Ilm. Teknol. Elektro*, vol. 20, no. 2, p. 195, Dec. 2021, doi: 10.24843/mite.2021.v20i02.p03.

[2] D. Anggarwati, O. Nurdiawan, I. Ali, and D. A. Kurnia, "Penerapan Algoritma K-Means Dalam Prediksi Penjualan Karoseri," vol. 1, no. 2, pp. 58–62, 2021.

[3] S. Butsianto and N. T. Mayangwulan, "Penerapan Data Mining Untuk Prediksi Penjualan Mobil Menggunakan Metode K-Means Clustering," 2020.

[4] D. Ayu, I. C. Dewi, and K. Pramita, "Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," 2019.

[5] Qomariyah and M. U. Siregar, "Comparative Study of K-Means Clustering Algorithm and K-Medoids Clustering in Student Data Clustering," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 2, pp. 91–99, 2022, doi: 10.14421/jiska.2022.7.2.91-99.

[6] A. Rizal, D. Candra, R. Novitasari, and M. Hafiyusholeh, "Pengelompokkan Karyawan Berdasarkan Kesalehan Menggunakan Perbandingan Fuzzy C-Means , K-Means , dan Probabilistic Distance Clustering," vol. 11, no. 2, pp. 69–77, 2022, doi: 10.14421/fourier.2022.112.69-77.

[7] P. R. N. Saputra and A. Chusyairi, "Perbandingan Metode Clustering dalam Pengelompokan Data Puskesmas," *J. Resti (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, pp. 5–12, 2021.