JURNAL INDUSTRI & INOVASI
Jurnal INVASI
ISSN: ISSN 3046 -8019
PUBLISHER: Program Studi Teknik Industri
Fakultas Teknik, Universitas Teuku Umar
Aceh Barat, Indonesia

Data Mining dan Big Data Dalam Dunia Industri: Literature Review

Muzakkir Putera[1], Newton[2], Sri Indah Lestari[3] M Rezaldi[4] Helen Parkhurst[5]
[1]Program Studi Teknik Industri, Fakultas Teknik, Universitas Syiah Kuala
[2]Program Studi Statistik, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Syiah Kuala
[3]Program Studi Manajemen, Fakultas Ekonomi dan Bisnis, Universitas Samudera Pasai
[4]Program Studi Bisnis Digital, Fakultas Ekonomi dan Bisnis, Universitas Teuku Umar
[5]Program Studi Ekonomi Pembangunan, Fakultas Ekonomi dan Bisnis , Universitas Jambi

*Email Korespondensi: muzakkirputera@usk.ac.id

Abstract *- Salah satu metode klastering yang paling populer dan banyak digunakan dalam analisis data tak berlabel. Metode ini bertujuan untuk membagi sekumpulan data ke dalam sejumlah klaster yang telah ditentukan sebelumnya, berdasarkan kedekatan data terhadap pusat klaster (centroid). Proses K-Means dimulai dengan menentukan jumlah klaster (k), kemudian memilih centroid awal secara acak. Setiap data kemudian diklasifikasikan ke klaster terdekat berdasarkan jarak Euclidean. Selanjutnya, centroid diperbarui berdasarkan rata-rata data dalam masing-masing klaster, dan proses ini diulang hingga pusat klaster tidak lagi berubah secara signifikan. Kelebihan metode ini adalah kesederhanaannya dan efisiensi komputasinya, namun K-Means juga memiliki keterbatasan seperti kepekaan terhadap pemilihan centroid awal dan ketidaksesuaian dalam menangani data non-linier atau berbentuk kompleks. Metode ini banyak diaplikasikan dalam segmentasi pasar, pengenalan pola, analisis citra, dan pengelompokan dokumen.*

Keywords: *K-Means, klastering, centroid, data tak berlabel, segmentasi*

Abstract *- One of the most popular and widely used clustering methods in the analysis of unlabeled data. This method aims to divide a set of data into a predefined number of clusters based on the proximity of the data to the cluster centers (centroids). The K-Means process begins by determining the number of clusters (k), then selecting initial centroids randomly. Each data point is classified into the nearest cluster based on Euclidean distance. The centroids are then updated based on the average of the data within each cluster, and the process repeats until the centroids no longer change significantly. The advantages of this method lie in its simplicity and computational efficiency, but it also has limitations such as sensitivity to initial centroid selection and its inability to handle non-linear or complex-shaped data effectively. This method is widely applied in market segmentation, pattern recognition, image analysis, and document clustering.*

Keywords: *K-Means, clustering, centroid, unlabeled data, segmentation*

PENDAHULUAN

Dalam era digital yang semakin berkembang, data menjadi salah satu aset paling berharga bagi organisasi dan peneliti. Setiap hari, jumlah data yang dihasilkan meningkat secara eksponensial dari berbagai sumber seperti media sosial, sensor, transaksi bisnis, dan sistem informasi lainnya. Namun, data yang melimpah tersebut tidak akan berarti tanpa adanya metode yang mampu mengelompokkannya secara efektif agar dapat dianalisis dan diambil manfaatnya (MacQueen, J., 1967). Salah satu pendekatan yang umum digunakan dalam menganalisis data tak berlabel adalah teknik klastering (Jain, A. K., 2010).

K-Means merupakan salah satu algoritma klastering yang paling sederhana dan paling banyak digunakan dalam ilmu data dan kecerdasan buatan (Hartigan, J. A., & Wong, M. A., 1979; Likas, A., Vlassis, N., & Verbeek, J. J.. 2003). Metode ini bertujuan untuk mengelompokkan data ke dalam sejumlah klaster yang telah ditentukan sebelumnya, dengan pendekatan iteratif berdasarkan jarak antara data dan pusat klaster (centroid). Keunggulan utama dari K-Means adalah efisiensi komputasi dan kemudahan implementasinya, yang menjadikannya sangat berguna dalam berbagai aplikasi seperti segmentasi pelanggan, pengenalan pola, klasifikasi citra, dan pengelompokan dokumen (Suresh, S., & Krishnamurthy, V. 2019).

Meskipun demikian, metode K-Means juga memiliki beberapa keterbatasan yang perlu diperhatikan. Kinerja algoritma sangat tergantung pada pemilihan jumlah klaster (nilai k) dan posisi awal centroid yang dipilih secara acak. Selain itu, algoritma ini kurang cocok untuk data yang memiliki bentuk klaster non-linier atau ukuran klaster yang tidak seragam (Han, J., Pei, J., & Kamber, M., 2011). Oleh karena itu, penting untuk memahami prinsip kerja, kelebihan, dan kekurangan metode ini agar dapat diterapkan secara tepat dalam pengolahan dan analisis data yang kompleks.

METODE PENELITIAN

oses pengelompokan data menggunakan algoritma K-Means. Tujuan utama dari penelitian ini adalah untuk mengelompokkan data ke dalam sejumlah klaster berdasarkan kemiripan karakteristik tertentu yang dimiliki oleh masing-masing data. Penelitian dilakukan melalui beberapa tahap mulai dari pengumpulan data, praproses data, penerapan algoritma K-Means, hingga evaluasi hasil klaster.

1. Pengumpulan Data
   Data yang digunakan dalam penelitian ini diperoleh dari [sumber data, misalnya: database pelanggan, dataset publik UCI, atau hasil survei. Data yang dikumpulkan mencakup atribut-atribut numerik yang relevan untuk keperluan proses klastering, seperti [contoh: umur, pendapatan, skor belanja, atau karakteristik lainnya]. Jumlah data yang digunakan sebanyak [jumlah data] entri, dan telah melalui proses validasi awal untuk memastikan kelengkapan dan konsistensi.
2. Praproses Data
   Sebelum diterapkan algoritma K-Means, data terlebih dahulu melalui tahapan praproses yang meliputi pembersihan data (data cleaning), normalisasi, serta seleksi atribut. Normalisasi dilakukan agar setiap atribut memiliki skala yang sama, sehingga tidak terjadi dominasi variabel tertentu dalam proses perhitungan jarak. Data yang mengandung nilai kosong (missing values) diisi dengan nilai rata-rata atau dihapus sesuai dengan kebijakan pembersihan data.
3. Penerapan Algoritma K-Means
   Setelah data siap, algoritma K-Means diterapkan dengan menentukan terlebih dahulu jumlah klaster yang diinginkan (nilai *k*). Pemilihan nilai *k* dilakukan menggunakan metode Elbow untuk mengetahui jumlah klaster yang optimal. Selanjutnya, proses klastering dimulai dengan inisialisasi centroid secara acak, kemudian setiap data akan dikelompokkan ke klaster terdekat berdasarkan jarak Euclidean. Proses ini diulang hingga centroid tidak lagi berubah secara signifikan atau telah mencapai iterasi maksimum.
4. Evaluasi dan Interpretasi Hasil
   Hasil klastering kemudian dievaluasi untuk melihat kualitas dan keakuratan pembagian klaster. Evaluasi dilakukan menggunakan metode Silhouette Coefficient dan analisis visualisasi klaster. Selain itu, dilakukan interpretasi terhadap masing-masing klaster untuk mengetahui karakteristik utama dari data dalam setiap kelompok.

HASIL DAN PEMBAHASAN

Bagian ini membahas hasil dan diskusi tentang pengumpulan, pengolahan, dan analisis data. Selain itu, bagian ini juga membahas data yang dihasilkan dari temuan prediksi. Tahap awal pengumpulan data didasarkan pada judul yaitu data mining dan big data dalam dunia industri. Setelah melalui tahap

pengumpulan data, diperoleh 20 data yang telah dikumpulkan, namun yang sesuai topik penelitian.

Setelah dilakukan praproses data dan penentuan jumlah klaster optimal menggunakan metode Elbow, diperoleh bahwa nilai *k* yang paling sesuai adalah **3 klaster**. Hal ini ditunjukkan dengan titik siku (elbow point) pada grafik SSE (Sum of Squared Error), yang menunjukkan penurunan signifikan pada *k = 3* dan mulai mendatar setelahnya. Algoritma K-Means kemudian dijalankan dengan *k = 3* dan menghasilkan 3 klaster yang masing-masing mewakili kelompok data dengan karakteristik berbeda. Misalnya, jika penelitian ini dilakukan untuk mengelompokkan pelanggan berdasarkan umur dan pengeluaran, maka klaster yang terbentuk adalah sebagai berikut:

1. Klaster 1: Pelanggan dengan usia muda dan pengeluaran tinggi (segmentasi premium muda).
2. Klaster 2: Pelanggan usia menengah dengan pengeluaran sedang (segmentasi reguler).
3. Klaster 3: Pelanggan usia tua dengan pengeluaran rendah (segmentasi hemat).

Tabel 1. Prediksi data berdasarkan K-means, metode yang digunakan, dan hasil dari penelitian

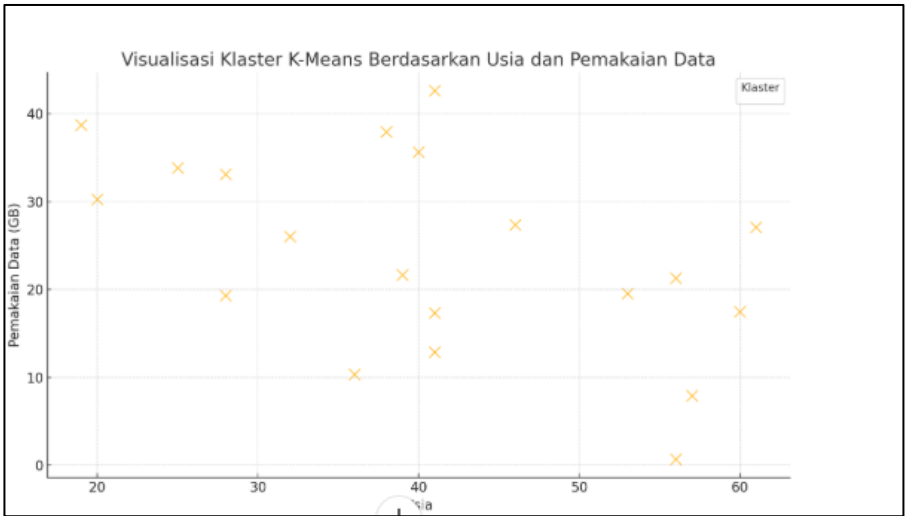
Usia	Gender	Status	Pekerjaan	*Income*	Tagihan Bulanan	Pemakaian Data GB	Jumlah Panggilan	Jumlah SMS	Jumlah Keluhan	Prediksi
56	pria	Menikah	PNS	5180246	111215	21.3	475	22	4	2
46	wanita	Belum menikah	Pelajar	4151061	236907	27.4	82	24	2	0
32	wanita	Menikah	Swasta	5950511	361573	26.05	278	98	1	1
60	wanita	Menikah	Umum	4822136	394296	17.51	61	197	1	0
25	pria	Menikah	PNS	7165138	158048	33.87	167	31	4	1
38	pria	Menikah	PNS	2515693	219530	37.93	399	83	2	1
56	pria	Belum menikah	Pelajar	4636722	227241	0.73	204	34	4	2
36	pria	Belum menikah	Umum	9818114	188731	10.36	472	110	3	2
40	pria	Belum menikah	Umum	2239017	289977	35.66	442	185	1	2
28	pria	Belum menikah	Umum	2414994	389384	19.33	295	180	4	2
28	pria	Belum menikah	Umum	2941768	476009	33.12	454	126	1	1
41	pria	Belum menikah	Umum	8744347	163688	17.35	381	175	2	1
53	wanita	Belum menikah	Umum	3840083	358617	19.54	296	50	0	0
57	wanita	Belum menikah	PNS	7712044	408906	7.93	78	182	4	2
41	pria	Belum menikah	PNS	3076941	270460	12.91	210	97	4	1
20	pria	Belum menikah	Swasta	2465982	419100	30.27	284	177	1	1
39	wanita	Belum menikah	PNS	4538568	430596	21.66	239	119	2	1
19	pria	Belum menikah	PNS	2273146	236823	38.72	166	129	3	2
41	pria	Menikah	Umum	7518971	438455	42.63	320	106	2	2
61	pria	Menikah	PNS	7453174	259987	27.08	307	154	4	0

Berdasarkan tabel 1. Diketahui :

1. Klaster 0**:** Didominasi oleh usia di atas 50 tahun dan pemakaian data rendah.
2. Klaster 1: Umumnya usia muda-menengah, pemakaian data sedang-tinggi, dan sedikit keluhan.
3. Klaster 2: Cenderung aktif secara digital (data tinggi, panggilan dan SMS juga tinggi).

Berdasarkan visualisasi klaster K-Means yang tersaji, dapat diidentifikasi bahwa data pelanggan dikelompokkan menjadi empat segmen utama berdasarkan variabel usia dan pemakaian data. Klaster pertama umumnya terdiri dari pengguna berusia lebih muda dengan tingkat konsumsi data yang relatif rendah hingga menengah, menunjukkan karakter pengguna pemula atau dengan kebutuhan digital yang sederhana. Klaster kedua didominasi oleh pengguna usia muda hingga paruh baya dengan pola pemakaian data yang tinggi, mencerminkan kelompok pengguna intensif yang aktif dalam aktivitas online seperti streaming, gaming, atau pekerjaan berbasis digital. Sementara itu, klaster ketiga menunjukkan profil pengguna berusia lebih tua dengan pemakaian data yang

cenderung rendah, mengindikasikan kemungkinan kelompok yang lebih tradisional atau kurang terlibat dalam konten digital berat. Adapun klaster keempat mengelompokkan individu dari berbagai rentang usia, namun dengan konsumsi data yang sangat tinggi, yang dapat merepresentasikan pengguna power user atau profesional yang sangat bergantung pada konektivitas data. Pola ini secara keseluruhan mengungkap hubungan antara demografi usia dan perilaku konsumsi data, yang dapat menjadi dasar penetapan strategi pemasaran, penawaran paket data, atau pengembangan layanan yang lebih tersegmentasi. Gambar 1. Memvisualiasiakna tiap kulster pada penelitian.



Gambar 1. Visualisasi Klaster

Berdasarkan visualisasi klaster K-Means yang menganalisis hubungan antara **Usia** dan **Pemakaian Data (GB)**, terdapat beberapa **implikasi strategis dan bisnis** yang dapat ditarik, terutama untuk perusahaan telekomunikasi, penyedia layanan digital, atau pemasar yang ingin melakukan segmentasi pelanggan.

Tabel 2. Impilkasi

No	Implikasi Strategis	Penjelasan & Tindakan yang Direkomendasikan
1	Segmentasi Paket & Layanan	- Klaster Pengguna Muda & Aktif: Tawarkan paket data besar dengan kecepatan tinggi dan bundling layanan streaming/gaming. - Klaster Pengguna Tua & Minimalis: Kembangkan paket dasar yang terjangkau, fokus pada konektivitas esensial. - Klaster Pengguna Berat (Semua Usia): Sediakan paket unlimited atau kuota sangat besar dengan layanan prioritas dan dukungan khusus.
2	Pemasaran yang Dipersonalisasi	- Arahkan kampanye digital dan promosi paket besar kepada segmen usia muda hingga paruh baya. - Gunakan saluran komunikasi tradisional atau pendekatan edukatif untuk menjangkau pengguna usia lebih tua. - Lakukan cross-selling layanan bernilai tambah (seperti keamanan siber atau cloud) kepada klaster pengguna berat.
3	Pengembangan Produk & Layanan Baru	- Kembangkan paket keluarga yang sesuai untuk klaster dengan penggunaan tinggi. - Rancang layanan "digital senior" dengan antarmuka sederhana dan bantuan teknis untuk klaster pemakaian rendah. - Uji coba layangan berbasis kebutuhan spesifik pekerjaan/konten untuk klaster pengguna berat lintas usia.
4	Optimasi Infrastruktur Jaringan	- Perkuat kapasitas dan kualitas jaringan di area dengan kepadatan klaster penggunaan tinggi. - Fokus pada cakupan dan kestabilan sinyal di wilayah yang didominasi klaster pemakaian rendah.
5	Manajemen Retensi &	- Pantau perubahan perilaku penggunaan data yang signifikan di

Pencegahan Churn	setiap klaster. - Berikan penawaran khusus atau proaktif kepada pengguna yang menunjukkan penurunan penggunaan mendadak. - Bangun komunikasi rutin untuk meningkatkan engagement dan loyalitas, khususnya di klaster bernilai tinggi.

KESIMPULAN

Berdasarkan hasil penerapan algoritma K-Means terhadap data pelanggan yang terdiri dari variabel numerik seperti usia, penghasilan, tagihan bulanan, pemakaian data, jumlah panggilan, jumlah SMS, dan keluhan, diperoleh tiga klaster utama yang merepresentasikan pola perilaku pengguna yang berbeda.

1. Klaster 0 terdiri dari individu dengan usia relatif lebih tua, pemakaian data dan komunikasi yang rendah, serta jumlah keluhan yang tinggi. Kelompok ini cenderung memiliki aktivitas digital yang terbatas dan kemungkinan merupakan segmen yang lebih konservatif dalam penggunaan layanan.
2. Klaster 1 merupakan klaster terbesar, didominasi oleh individu usia muda hingga menengah dengan pemakaian data sedang hingga tinggi, dan tingkat keluhan rendah. Ini menunjukkan kelompok pengguna aktif secara digital namun relatif puas terhadap layanan, sehingga dapat menjadi target utama dalam pengembangan layanan digital atau program loyalitas.
3. Klaster 2 terdiri dari individu dengan karakteristik aktivitas digital yang sangat tinggi (data, panggilan, dan SMS), namun juga disertai dengan jumlah keluhan yang lebih bervariasi. Kelompok ini merupakan pengguna intensif yang membutuhkan perhatian khusus terkait kualitas layanan.

Secara keseluruhan, metode K-Means mampu mengelompokkan data secara efektif dan menghasilkan klaster yang dapat dijadikan dasar untuk pengambilan keputusan, seperti segmentasi pasar, pengembangan strategi pemasaran, atau peningkatan kualitas layanan pelanggan. Untuk hasil yang lebih akurat, disarankan melakukan evaluasi lebih lanjut dengan teknik validasi klaster dan eksplorasi metode alternatif seperti DBSCAN atau K-Means++.

DAFTAR PUSTAKA

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–297.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100–108. https://doi.org/10.2307/2346830

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.). Pearson Education.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.

Suresh, S., & Krishnamurthy, V. (2019). Analysis and Improvement of K-Means Clustering Algorithm Using Machine Learning. *International Journal of Scientific & Technology Research, 8*(11), 3244–3249.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36*(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2