EXPLORE
Jurnal Sistem Informasi dan Telematika
(Telekomunikasi, Multimedia dan Informatika)

Komparasi Algoritma Hierarchical, K-Means, dan DBSCAN pada Analisis Data Penjualan Melalui Facebook

Farah Dwi Wahyuningtyas, Abdillah Arafat, Agus Stiawan, Dwi Rolliawati
Prodi Sistem Informasi, Fakultas Sains dan Teknologi
Universitas Islam Negeri Sunan Ampel Surabaya
Surabaya, Indonesia
h76219021@student.uinsby.ac.id, h06219001@student.uinsby.ac.id,
h96219038@student.uinsby.ac.id, dwi_roll@uinsby.ac.id

Abstract-The use of the internet in Indonesia to access social media has increased from the previous four years, where 36.36% of users still use social media Facebook. The average social media users are teenagers with smartphones. Facebook has features that are favored by its users for buying and selling activities, so that it can increase user engagement and sales data. To analyze the increase in sales data, this study uses data mining with clustering methods. By using secondary data from the UCI Repository, a comparative analysis of three different algorithms was carried out to find out which is the best among the Hierarchical, K-Means, and DBSCAN algorithms. The results showed that the Hierarchical algorithm obtained the highest silhouette score, namely 0.884, a fairly thin difference with the silhouette score obtained by K-Means, which was 0.872. Furthermore, the results of comparisons made using performance indicators show that K-Means is the best algorithm with an average execution time of 0.402 seconds, a considerable difference from the other two algorithms. Based on the two indicators that have been used, it can be seen that the best algorithm for analyzing sales data via Facebook is the K-Means algorithm. Finally, the appearance of the number of clusters 2 from the K-Means algorithm can group sales data via Facebook into two categories, namely "Busy Posts" and "Lone Posts".

Keywords: Facebook, Clustering, Hierarchical, K-Means, DBSCAN

Abstrak-Penggunaan internet di Indonesia untuk akses media sosial meningkat dari empat tahun sebelumnya, di mana 36,36% pengguna masih menggunakan media sosial Facebook. Rata-rata pengguna media sosial ini berasal dari kalangan remaja dengan smartphone. Facebook memiliki fitur-fitur yang digemari oleh penggunanya untuk melakukan aktivitas jual beli, sehingga dapat meningkatkan user engagement dan data penjualan. Untuk menganalisis peningkatan data penjualan, penelitian ini menggunakan data mining dengan metode klasterisasi. Dengan menggunakan data sekunder dari UCI Repository, dilakukan analisis terhadap komparasi tiga algoritma berbeda untuk mengetahui mana yang terbaik di antara algoritma Hierarchical, K-Means, dan DBSCAN. Hasil penelitian menunjukkan bahwa algoritma Hierarchical dengan memperoleh skor silhouette tertinggi yaitu 0.884, selisih yang cukup tipis dengan perolehan silhouette score yang diperoleh K-Means sebesar 0.872. Selanjutnya, hasil komparasi yang dilakukan dengan menggunakan indikator performa menunjukkan bahwa K-Means merupakan algoritma terbaik dengan rata-rata waktu eksekusi selama 0.402 detik, selisih yang cukup jauh dari dua algoritma yang lain. Berdasarkan dua indikator yang telah digunakan tersebut, dapat diketahui bahwa algoritma terbaik untuk menganalisis data penjualan melalui Facebook adalah algoritma K-Means. Terakhir, munculnya jumlah cluster 2 dari algoritma K-Means dapat mengelompokkan data penjualan melalui Facebook menjadi dua kategori, yaitu "Postingan Ramai" dan "Postingan Kurang Ramai".

Kata Kunci: Facebook, Klasterisasi, Hierarchical, K-Means, DBSCAN

1. Pendahuluan

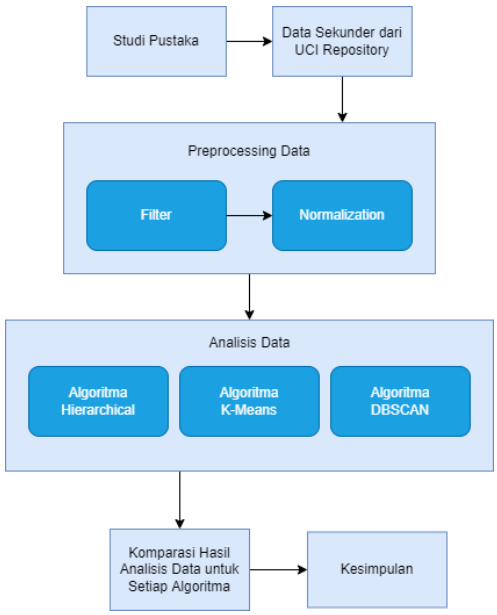
Survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menghasilkan fakta bahwa pengguna internet baru di Indonesia telah mengalami kenaikan sebesar 25,5 Juta, setara dengan 8,9% jika dibandingkan dengan 2018 silam yang total jumlah penggunanya mencapai 63 juta orang [1]. Dari angka tersebut, diketahui 95% pengguna internet mengakses media sosial. Semakin banyaknya media sosial yang tersedia di zaman modern ini, 36,36% pengguna internet masih menggunakan Facebook untuk saling berinteraksi dengan pengguna lain [2]. Jumlah tersebut telah mengantarkan Indonesia ke posisi nomor 4 dunia dengan pengguna Facebook terbanyak setelah USA, Brazil, dan India [3]. Penggunaan Facebook di zaman sekarang tidak lagi hanya terbatas pada konsep just fun, tetapi sangat bisa juga digunakan sebagai salah satu media penjualan secara online yang lebih lumrah dikenal dengan sebutan e-commerce, yaitu sarana pemasaran secara elektronik [4]. Tidak hanya menjadi sosial media sekaligus marketplace yang hanya dapat menampilkan produk saja,

tetapi para penjual di Facebook juga bisa memanfaatkan fitur live streaming untuk mempertunjukkan produk yang mereka miliki secara interaktif, mempromosikan produk/perusahaan mereka, dan melakukan tutorial pemakaian dari produk-produk tertentu [5]. Setyawan, dkk membuktikan bahwa rata-rata pelajar yang mempunyai smartphone mengakses lebih dari 6 jam per hari untuk aplikasi yang dilengkapi dengan fitur live streaming. Tidak heran jika para pengguna media sosial dengan fitur live streaming telah mendapatkan peningkatan dalam user engagement yang disebabkan oleh adanya interaksi secara langsung dan komunikasi dua arah dalam kegiatan live streaming tersebut [6]. Berdasarkan peningkatan data penjualan melalui Facebook tersebut, akan dilakukan analisis menggunakan salah satu peran dari data mining, yaitu clustering. Dalam metode clustering itu sendiri, ada beberapa algoritma yang bisa diterapkan, di antaranya adalah Hierarchical, K-Means, dan DBSCAN. Dari beberapa algoritma yang telah disebutkan, penting dilakukan analisis guna mengetahui algoritma terbaik yang bisa diterapkan untuk analisis data penjualan melalui Facebook. Terdapat beberapa penelitian terdahulu yang telah melakukan perbandingan untuk algoritma K-Means dan DBSCAN. Hasil penelitian Sisca dkk (2016) menunjukkan bahwa metode K-Means lebih baik dari metode DBSCAN dalam mengelompokkan data rumah kost [7]. Hal ini berbanding terbalik dengan penelitian terbaru yang dilakukan oleh Mustika dkk (2021) untuk pengelompokan status desa dan penelitian yang dilakukan oleh Rimelda dkk (2021) untuk pengelompokan kasus Covid-19 [8][9]. Hasil dari kedua penelitian tersebut justru menunjukkan bahwa metode DBSCAN lebih baik dari metode K-Means. Ketiga penelitian terdahulu ini hanya melakukan perbandingan terhadap algoritma DBSCAN dan K-Means, oleh karena itu penelitian ini diharapkan dapat memberikan kontribusi keterbaruan dengan melakukan komparasi terhadap algoritma Hierarchical.

2. Metodologi

Metodologi yang digunakan untuk melakukan penelitian ini adalah dimulai dengan melakukan studi pustaka dan pencarian data sekunder melalui UCI Repository. Kemudian, data tersebut perlu dilakukan preprocessing terlebih dahulu sebelum dilakukan analisis menggunakan algoritma Hierarchical, K-Means, dan DBSCAN. Setelah itu, dilakukan perbandingan untuk hasil analisis dari masing-masing algoritma tersebut. Terakhir, dapat ditarik kesimpulan terkait algoritma mana yang terbaik untuk diterapkan dalam analisis data penjualan melalui live Facebook.

Langkah-langkah yang dilakukan tersebut akan dijelaskan lebih lanjut pada subbab maupun bab terkait pada bagian berikutnya. Adapun gambar terkait metodologi di atas adalah sebagai berikut.



Gambar 1. Bagan Alir Metodologi

A. Studi Pustaka

Sebelum melakukan penelitian, perlu dilakukan sebuah studi pustaka. Penelitian ini menggunakan teori dasar dan memanfaatkan teknik penggalian data yang disebut data mining. Data mining merupakan proses identifikasi pola pada data dengan tujuan mendapatkan informasi berguna dari data yang tersebar dan bersifat

besar. Data mining digunakan untuk menentukan tujuan hingga penilaian [10]. Adapun teknik dalam melakukan penggalian data yaitu klasterisasi atau clustering. Clustering merupakan pengelompokan data serupa atau mirip dengan mengelompokkan data menjadi beberapa kelompok (cluster) [11]. Clustering melalui proses partisi data menjadi himpunan-himpunan sesuai objek dataset. Klasterisasi dinilai membantu untuk menganalisis dengan identifikasi objek, hal ini disebut segmentasi data. Setelah dilakukan pengelompokkan data dan mendapatkan hasil dari tiap setiap algoritma, dilakukan studi komparasi untuk membandingkan dua hingga lebih objek penelitian. Hal ini dilakukan untuk menemukan persamaan dan perbedaan dari objek-objek tersebut [12].

B. Data

Data yang digunakan pada penelitian ini merupakan data sekunder. Menurut Edi Riadi (2016), data sekunder adalah jenis data yang didapatkan secara tidak langsung dari objek penelitian [13]. Adapun data yang digunakan adalah dataset live sellers in Thailand atau data pengguna untuk melakukan kegiatan jual beli secara online yang diperoleh dari sumber data terbuka UCI Repository [14].Dalam data berjumlah 7050 tersebut, terdapat 12 fitur atau parameter yang ada di dalamnya. Adapun penjelasan terkait parameter tersebut dapat dilihat pada Tabel 1.

Tabel 1 Parameter Dataset

No	Parameter	Deskripsi
1	status_id	Nomor urutan data
2	status_type	Jenis postingan yang dibuat oleh seller, *photo* atau *video*
3	status_published	Tanggal postingan dibuat
4	num_reactions	Jumlah akun yang memberikan *reaction*
5	num_comments	Jumlah akun yang memberikan komentar
6	num_shares	Jumlah akun yang membagikan postingan
7	num_likes	Jumlah akun yang menyukai postingan
8	num_loves	Jumlah akun yang memberikan simbol *love*
9	num_wows	Jumlah akun yang memberikan simbol wow
10	num_hahas	Jumlah akun yang memberikan simbol haha
11	num_sads	Jumlah akun yang memberikan simbol *sad*
12	num_angrys	Jumlah akun yang memberikan simbol *angry*

C. Preprocessing Data

Preprocessing data merupakan salah satu tahap dari serangkaian data mining yang meliputi persiapan dan transformasi data menjadi bentuk yang sesuai dengan prosedur data mining. Preprocessing data bertujuan untuk memperkecil ukuran data, menemukan relasi antar data, menormalkan data, menghapus outlier, dan mengekstrak fitur untuk data. Preprocessing data mencakup beberapa teknik seperti pembersihan data, integrasi, transformasi, dan pengurangan [15]. Adapun tahapan preprocessing data dalam proses clustering dapat dilihat pada gambar di bawah ini. Proses preprocessing data sangat penting untuk dilakukan guna mengubah sumber data menjadi format yang sesuai dan mudah untuk dilakukan proses pengklasteran sehingga proses clustering tersebut dapat lebih optimal [16].

Perlakuan preprocessing data yang dilakukan dalam penelitian ini adalah filter dan normalization. Filter yang dimaksud di sini bukan untuk memilih value tertentu dari sebuah kolom, melainkan untuk memilih beberapa atribut data yang akan diproses. Adapun data yang digunakan adalah semua data yang tercantum dalam Tabel 1 selain status_id, status_type, dan status_published. Data tersebut tidak dapat digunakan dalam proses clustering karena termasuk ke dalam data kategorikal. Preprocessing data yang dilakukan selanjutnya adalah normalization. Normalisasi data diperlukan ketika berurusan dengan atribut pada skala yang berbeda. Jika tidak dilakukan normalisasi, maka dapat menyebabkan dilusi efektivitas atribut penting yang sama pentingnya (pada skala yang lebih rendah) karena atribut lain memiliki nilai pada skala yang lebih besar. Dengan kata lain, ketika ada banyak atribut dengan skala yang berbeda, ini dapat menyebabkan model data yang buruk ketika proses data mining. Jadi, diperlukan normalisasi data untuk semua atribut pada skala yang sama. Normalisasi adalah teknik penskalaan atau teknik pemetaan atau tahap pra pemrosesan, di mana dapat ditemukannya rentang baru dari rentang yang sudah ada [17].

Adapun rumus untuk masing-masing perlakuan pada preprocessing data yaitu Standardization, Normalization, dan Min-max scaler adalah sebagai berikut [18]:

Standardization:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

with mean:

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i1.2931

$$\mu = \frac{1}{N}\sum_{i=1}^{N} (x_i) \quad (2)$$

And standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2} \quad (3)$$

Normalization

L1: $z = ||x||_1 = \sum_{i=1}^{n} |x_i|$ (4)

L2: $z = ||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ (5)

Max: $z = ||x||_\infty = max\,|x_i|$ (6)

Min max scaler

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (7)$$

D. Algoritma

Algoritma merupakan sebuah sistem kerja komputer yang mencakup software, hardware, dan brainware. Tanpa adanya salah satu dari ketiga komponen tersebut, komputer tidak akan dapat memberikan manfaat apapun. Kita hanya akan terpaku pada software yang kita gunakan. Sedangkan software terbangun atas susunan program dan syntax (cara penulisan/pembuatan program). Dalam menyusun program atau syntax itulah, diperlukan langkah-langkah yang sistematis dan logis agar dapat memecahkan suatu masalah atau tujuan tertentu dalam proses rancang bangun sebuah software. Algoritma mempunyai peran yang sangat penting dalam penyusunan program atau syntax tersebut. Pengertian dari algoritma itu sendiri adalah sebuah susunan logis dan sistematis yang digunakan untuk menyelesaikan suatu masalah tertentu atau untuk mencapai sebuah maksud dan tujuan. Dalam dunia komputer, algoritma mempunyai peran yang sangat penting pada rancang bangun sebuah software. Dalam kehidupan sehari-hari, tidak dapat kita pungkiri bahwa algoritma telah masuk dalam semua lini kehidupan kita [19].

Beberapa algoritma yang dilakukan dalam penelitian ini adalah Hierarchical, K-Means dan DBSCAN. Adapun penjelasan dari masing-masing algoritma tersebut adalah sebagai berikut.

1. Hierarchical

Hierarchical merupakan salah satu algoritma dalam metode *clustering* yang bisa digunakan untuk mengelompokkan dokumen *(document clustering)*. Dengan algoritma ini, bisa diperoleh sebuah kumpulan partisi yang berurutan, dimulai dari beberapa *cluster* yang berada di tingkatan paling bawah hingga *single cluster* yang berada di tingkatan paling atas. *Cluster-cluster* yang berada di tingkatan paling bawah adalah kumpulan *cluster* yang mempunyai unsur-unsur individu, sedangkan *single cluster* yang berada di tingkatan paling atas adalah sebuah *cluster* yang di dalamnya mengandung unsur yang dimiliki oleh keseluruhan *cluster* [20].

Beberapa metode dalam algoritma *Hierarchical* yang sering digunakan adalah *Single Linkage, Complete Linkage, Average Linkage, Average Group Linkage,* dan masih banyak lagi. Layaknya *partition-based clustering,* jarak bisa dipakai untuk menghitung tingkat kemiripan yang dimiliki oleh antar data [21].

Algoritma *Hierarchical* bisa direpresentasikan dalam bentuk visual melalui dendogram. Dendogram disusun dengan membuat *similarity matrix* yang dapat mengelompokkan tingkatan dari kemiripan antar data. Tingkat kemiripan ini dapat dihitung menggunakan beberapa cara, seperti *Euclidean Distance Space* dan *Manhattan Distance*. Adapun formula dari kedua cara tersebut adalah sebagai berikut.

Manhattan Distance

$$D_{man}(x, y) = \sum_{j=1}^{d} |x_j - y_j| \quad (8)$$

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i1.2931

Euclidean Distance

$$D_{(x_2,x_1)} = \sqrt{\sum_{j=1}^{d} \left|x_{2j} - x_{1j}\right|^2} \quad (9)$$

2. K-Means

Algoritma ini paling banyak digunakan dan sering dijumpai pada kasus klasterisasi. *K-Means* membagi titik *M* ke dalam dimensi *N* menjadi *K*, untuk meminimalkan kuadrat pada klaster [22]. *K-Means* termasuk ke dalam *unsupervised learning*, yaitu pengelompokkan pola-pola data tanpa terpengaruh jumlah *cluster*. *Dataset* yang digunakan memiliki fungsi X = $\{x_1, \ldots, x_n\}$[23]. Kemudian *K-Means* meminimalkan fungsi objek $\sum_i^n = 1 \sum_k^c = 1 z_{ik} \left|\left|x_i - a_k\right|\right|^2$ (10)

Berikut merupakan tahapan untuk *clustering* dengan K-Means.

1) Menentukan nilai *k* (jumlah klaster)
2) Menentukan *k center* dari *dataset*
3) Menggunakan rumus jarak *Euclidean* dari tiap *centroid,* yang mana *x* dan *y* sebagai koordinat objek, *s* dan *t* sebagai koordinat *centroid* :
$$De = \sqrt{(x_i - s_i)^2 (y_i - t_i)^2} \quad (11)$$
4) Menentukan jarak terdekat berdasarkan letak *centroid*
5) Menentukan rata-rata nilai *centroid* yang baru dengan rumus: $v_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj}$ (12)
di mana rata-rata dari cluster *i* untuk variabel *j,* N_i jumlah data, *i* dan *k* menjadi indeks klaster, *j* merupakan indeks variabel, X_{kj} nilai data *k* untuk variabel *j*.
6) Kelima tahapan tersebut dilakukan hingga klaster tidak berubah.

3. DBSCAN

DBSCAN *(Density-Based Spatial Clustering of Application with Noise)* merupakan salah satu algoritma yang menjadi pelopor dalam perkembangan metode *clustering* dengan berdasarkan pada kepadatan atau yang lumrah dikenal dengan istilah *density based clustering* dalam dunia *data mining* [24].

Metode yang digunakan dalam algoritma DBSCAN adalah dengan cara membatasi wilayah tertentu berdasarkan kepadatan yang saling terhubung satu sama lain *(density-connected)*. Setiap objek dari sebuah cakupan wilayah *(cluster)* harus terdapat setidaknya sejumlah minimum data. Semua objek yang bukan merupakan bagian dari *cluster* tertentu dianggap sebagai sebuah *noise*. Adapun langkah-langkah perhitungan dalam algoritma DBSCAN ini adalah sebagai berikut.

a. Inisialisasi parameter *mintps, eps*
b. Menentukan titik awal atau *p* dengan cara acak
c. Mengulangi langkah 1-3 untuk semua titik
d. Menghitung *eps* atau semua jarak antar titik yang kepadatannya dapat dijangkau *(density reachable)* terhadap *p* menggunakan rumus:
$$E(x,y) = \sqrt{\sum_{i=0}^{n} (X_i - Y_i)^2} \quad (13)$$
e. Jika titik yang memenuhi *eps* ternyata lebih besar dari *mintps*, maka titik *p* dijadikan sebagai *core point* dari *cluster* yang telah terbentuk
f. Namun jika *p* ternyata adalah *border point* dan tidak ada titik yang *density reachable* terhadap *p,* maka proses dilanjutkan dengan cara melakukan perhitungan terhadap titik yang lain.

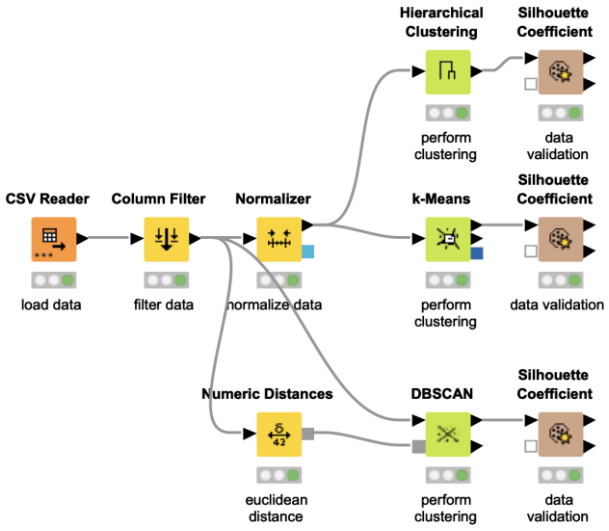
3. Hasil dan Pembahasan

Proses analisis komparasi algoritma dilakukan dengan menggunakan software KNIME dan bahasa pemrograman Python. KNIME (Konstanz Information Miner) merupakan sebuah platform yang dapat digunakan untuk mengintegrasikan, memproses, dan menganalisis data dari berbagai sumber. KNIME menawarkan berbagai alat visual yang memungkinkan pengguna membangun alur kerja analisis data langsung tanpa harus menuliskan baris kode. Python adalah bahasa pemrograman interpreted tingkat tinggi yang memiliki tujuan seperti bahasa pemrograman tingkat tinggi pada umumnya. Filosofi bahasa pemrograman Python desainnya menekankan keterbacaan kode dengan penggunaan indentasi yang signifikan. Konstruksi bahasa serta pendekatan berorientasi objek dari Python itu sendiri memiliki tujuan untuk membantu programmer menulis kode yang jelas dan logis untuk proyek berskala kecil dan besar [25]. Proses analisis dilakukan dengan memodelkan setiap metode clustering untuk memberikan gambaran alur yang nantinya akan dieksekusi menggunakan bahasa pemrograman Python.

Gambar 2 merupakan tampilan dari workflow pada KNIME yang diawali dengan membaca data menggunakan node CSV Reader. Sebelum mulai diproses, perlu dilakukan tahap preprocessing data terlebih dahulu menggunakan node Column Filter untuk memfilter parameter yang akan digunakan dan dilanjutkan dengan normalisasi menggunakan node Normalizer. Untuk algoritma DBSCAN sendiri, ada satu node tambahan yang diperlukan sebelum memproses data, yaitu Numeric Distances untuk menentukan jarak Euclidean dari data numerik yang ada. Selanjutnya, dilakukan analisis menggunakan tiga node berbeda sesuai dengan algoritma terkait. Dari masing-masing algoritma

tersebut, diperlukan validasi menggunakan node Silhouette Coefficient.
Penjelasan dari masing-masing node yang digunakan pada workflow KNIME pada penelitian ini disajikan pada Tabel 2.

Setelah dilakukan visualisasi menggunakan KNIME, dilakukan analisis pada Python untuk mengetahui silhouette score yang dihasilkan dari setiap perlakuan yang diberikan pada masing-masing algoritma. Adapun hasil analisis yang dimaksud tertuang pada beberapa tabel 3.



Gambar 2. Model KNIME

Tabel 3. Silhouette Score Hierarchical

Preprocessing	*Cluster*								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Tanpa Scaler	0.8844	0.7729	0.7924	0.7937	0.7515	0.7512	0.6596	0.6697	0.6401
Standard Scaler	0.7961	0.7486	0.7533	0.7563	0.7571	0.7583	0.6588	0.6623	0.6637
Normalization	0.7749	0.6811	0.7052	0.6537	0.6476	0.6516	0.6618	0.4689	0.4636
Min Max Scaler	0.8042	0.8081	0.7495	0.6184	0.5732	0.5754	0.5807	0.5815	0.5832

Tabel 3 menunjukkan silhouette score untuk metode Hierarchical dari beberapa cluster, mulai dari cluster 2 hingga cluster 10. Adapun perolehan silhouette score tertinggi dari masing-masing perlakuan pre-processing data pada metode Hierarchical berturut-turut adalah 0.8844 untuk preprocessing tanpa scaler, 0.7961 untuk Standard scaler, 0.7749 untuk Normalization, dan 0.8042 untuk Min Max Scaler. Semua perolehan skor tertinggi terjadi pada cluster 2. Dari keseluruhan skor pada masing-masing preprocessing tersebut, skor tertinggi yang diperoleh pada metode Hierarchical adalah 0.8844 pada perlakuan tanpa scaler.

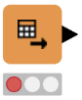
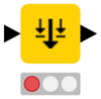
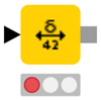
Tabel 4 Silhouette Score K-Means

Preprocessing	*Cluster*								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Tanpa Scaler	0.8722	0.8121	0.8176	0.7852	0.7600	0.7614	0.7614	0.6575	0.6819
Standard Scaler	0.8158	0.7500	0.7577	0.7605	0.7151	0.7118	0.7137	0.7108	0.6910

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i1.2931

Normalization	0.7998	0.8156	0.7479	0.7095	0.6220	0.6202	0.5595	0.5700	0.5488
Min Max Scaler	0.8125	0.8170	0.6443	0.6962	0.6196	0.6329	0.6359	0.6365	0.6319

Tabel 2 *Node* pada KNIME

Node	Nama	Fungsi
	CSV Reader	Untuk membaca file CSV
	Column Filter	Untuk memfilter kolom tertentu dari inputan data
	Normalizer	Untuk menormalkan semua nilai kolom numerik
	Numeric Distances	Untuk menentukan jarak Euclidean pada kolom numerik
	Hierarchical Clustering	Untuk mengelompokkan data inputan berdasarkan hirarki yang terbentuk
	k-Means	Untuk menentukan pusat klaster berdasarkan jumlah klaster yang telah ditentukan sebelumnya
	DBSCAN	Untuk menemukan klaster dalam database spasial besar dengan noise
	Silhouette Coefficient	Untuk menghitung koefisien silhouette berdasarkan jumlah klaster yang telah ditemukan

Tabel 4 menunjukkan silhouette score untuk metode K-Means dari beberapa cluster, mulai dari cluster 2 hingga cluster 10. Adapun perolehan silhouette score tertinggi dari masing-masing perlakuan preprocessing data pada metode K-Means berturut-turut adalah 0.8722 untuk preprocessing tanpa scaler, 0.8158 untuk Standard scaler, 0.8156 untuk Normalization, dan 0.8170 untuk Min Max Scaler. Perolehan skor tertinggi terjadi pada cluster 2 untuk perlakuan tanpa scaler dan standard scaler, sedangkan perolehan skor tertinggi pada Normalization dan Min Max Scaler terjadi pada cluster 3. Dari keseluruhan skor pada masing-masing preprocessing tersebut, skor tertinggi yang diperoleh pada metode K-Means adalah 0.8844 pada perlakuan tanpa scaler. Terakhir, sebelum dilakukan proses clustering menggunakan metode DBSCAN, dibutuhkan dua parameter input, yaitu Epsilon dan Minimum Points. Epsilon adalah jarak maksimal antara dua data dalam satu cluster yang memungkinkan, sedangkan minimum points adalah banyaknya data minimal dalam jarak epsilon agar terbentuk sebuah cluster. Adapun metode jarak yang digunakan dalam DBSCAN pada penelitian ini adalah jarak Euclidean. Angka-angka Epsilon yang digunakan pada penelitian ini dimulai dari 1.00 hingga 4.00 dengan menggunakan kelipatan 0.25. Sedangkan empat angka Minimum Points yang digunakan adalah 10, 15, 20, dan 25. Penggunaan angka-angka ini adalah dikarenakan dara yang dimiliki memiliki lebih dari 2 dimensi, sehingga MinPts = 2*dim, di mana redup = dimensi dari kumpulan data yang digunakan [26].

Tabel 6 Silhouette Score DBSCAN dengan Standard Scaler

Epsilon	*Minimum Points*							
	10		15		20		25	
	Output Cluster	*Silhouette Score*	*Output Cluster*	*Silhouette Score*	*Output Cluster*	*Silhouette Score*	*Output Cluster*	*Silhouette Score*
1.00	41	-0.2502	33	-0.2638	29	-0.2760	28	-0.2834
1.25	41	-0.2502	33	-0.2638	29	-0.2760	28	-0.2834
1.50	2	-0.4002	5	-0.4436	3	-0.1991	1	0
1.75	7	-0.3685	5	-0.4513	3	-0.1427	4	-0.1701
2.00	3	-0.0493	2	-0.0641	2	-0.1146	2	-0.1219
2.25	4	-0.2721	2	-0.0606	2	-0.1063	2	-0.1200
2.50	6	-0.0813	2	-0.0427	2	-0.0554	2	-0.3507
2.75	4	-0.1305	2	-0.0155	2	-0.0474	1	0
3.00	4	0.0255	2	0.0077	4	-0.1889	2	-0.0361
3.25	3	0.0723	2	0.0197	2	0.0010	2	-0.0304
3.50	5	0.0614	3	0.0351	1	0	3	-0.0093
3.75	5	-0.3142	3	0.0367	1	0	1	0
4.00	5	-0.3709	3	0.0888	1	0	1	0

Tabel 5 menunjukkan hasil silhouette score pada metode DBSCAN dengan perlakuan preprocessing Standard Scaler. Dapat diketahui bahwa perolehan silhouette score tertinggi pada nilai minimum points 10 adalah 0.6990, pada nilai minimum points 15 adalah 0.5156, pada nilai minimum points 20 adalah 0.5154, dan pada nilai minimum points 25 adalah 0.5145. Dari beberapa score tersebut, 0.6990 merupakan silhouette score tertinggi dari minimum points 10 dan epsilon 2.00 dengan jumlah cluster yang muncul sebanyak 2. Perlakuan preprocessing selanjutnya yang dilakukan pada metode DBSCAN adalah Normalizer dan Min Max Scaler. Dengan menggunakan beberapa epsilon dan minimum points yang sama dengan dua perlakuan sebelumnya, diketahui bahwa semua silhouette score yang muncul adalah 0. Terakhir, dari beberapa perlakuan preprocessing untuk metode DBSCAN, dilakukan komparasi untuk menentukan perolehan silhouette score tertinggi dari semua perlakuan yang telah diberikan. Adapun komparasi yang dimaksud pada gambar tabel 7.

Tabel 7 Silhouette Score DBSCAN

No	Perlakuan	*Cluster*	*Silhouette Score*
1	Tanpa Scaler	3	0.0887
2	Standard Scaler	2	0.6989
3	Normalizer	1	0
4	Min Max Scaler	1	0

Tabel 7 menunjukkan bahwa hasil metode DBSCAN dengan Standard Scaler dan dengan menggunakan 2 cluster pada studi kasus ini memiliki nilai Silhouette Score paling tinggi. Jenis scaler Normalizer dan Min Max Scaler tidak memiliki nilai Silhouette karena scaler tersebut hanya menghasilkan 1 cluster, sedangkan syarat validasi Silhouette Score adalah setidaknya memiliki 2 cluster. Selanjutnya, dilakukan kompresi dari semua metode yang telah dilakukan, yaitu Hierarchical, K-Means, dan DBSCAN berdasarkan silhouette score pada masing-masing metode tersebut. Adapun hasil komparasi tersebut pada tabel 8.

Tabel 8 Komparasi Hasil Silhouette Score

Metode	Perlakuan	*Cluster*	*Silhouette Score*

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i1.2931

Hierarchical	Tanpa Scaler	2	0.8843
K-Means	Tanpa Scaler	2	0.8721
DBSCAN	Standard Scaler	2	0.6989

Tabel 8 menunjukkan bahwa metode K-Means dan *Hierarchical* tanpa scaler dan dengan menggunakan 2 cluster memiliki nilai *Silhouette Score* cukup tinggi dan memiliki selisih yang tidak terlalu jauh. Selain berdasarkan *silhouette score*, perlu juga dilakukan komparasi untuk masing-masing metode dengan menggunakan indikator performa algoritma. Adapun hasil komparasi yang dimaksud pada tabel 9.

Tabel 9 Performa Algoritma

	Hierarchical	**K-Means**	**DBSCAN**
Rata-rata waktu eksekusi (detik)	1.6069	0.4016	0.8349

Performa algoritma pada Tabel 9 menunjukkan bahwa metode K-Means memiliki waktu eksekusi yang paling cepat. Metode DBSCAN memiliki waktu dua kali lipat lebih lama dan metode Hierarchical memiliki waktu empat kali lebih lama daripada eksekusi metode K-Means.

4. Kesimpulan

Hasil penelitian menunjukkan bahwa algoritma Hierarchical dengan perlakuan tanpa scaler dan jumlah cluster yang muncul sebanyak 2 memperoleh silhouette score tertinggi, yaitu 0.884. Perolehan silhouette score tertinggi ini disusul oleh angka 0.872 pada algoritma K-Means dengan perlakuan tanpa scaler dan jumlah cluster yang muncul sebanyak 2. Di posisi terakhir, ada algoritma DBSCAN yang memperoleh silhouette score 0.699 dengan jumlah cluster sebanyak 2 tetapi dengan jenis perlakuan yang berbeda dengan dua algoritma yang lain, yaitu Standard Scaler. Selanjutnya, hasil komparasi yang dilakukan dengan menggunakan indikator performa menunjukkan bahwa K-Means merupakan algoritma terbaik dengan rata-rata waktu eksekusi selama 0.402 detik. Di posisi kedua, ada DBSCAN yang memiliki rata-rata waktu eksekusi 0.835 detik, sekitar dua kali lipat dari performa K-Means. Angka ini kemudian disusul oleh performa dari algoritma Hierarchical yang rata-rata waktu eksekusinya adalah selama 1.607 detik. Berdasarkan dua indikator yang telah digunakan tersebut, dapat diketahui bahwa algoritma terbaik untuk menganalisis data penjualan melalui Facebook adalah algoritma K-Means. Hal ini dikarenakan jumlah silhouette score yang diperoleh oleh K-Means selisih cukup tipis dengan perolehan silhouette score tertinggi pada algoritma Hierarchical. Selain itu, performa yang dimiliki oleh K-Means jauh lebih baik dari kedua algoritma yang lain. Terakhir, munculnya jumlah cluster 2 dari algoritma K-Means dapat mengelompokkan data penjualan melalui Facebook menjadi dua kategori, yaitu "Postingan Ramai" dan "Postingan Kurang Ramai".

Saran untuk peneliti selanjutnya adalah dapat melakukan komparasi terhadap ketiga algoritma yang sama dengan dataset berbeda untuk membuktikan apakah K-Means memang merupakan algoritma andal untuk semua jenis dataset, atau hanya terbatas pada dataset tertentu saja.

5. Acknowledgements

Penulisan artikel ini tidak akan dapat selesai jika tidak adanya dukungan dari berbagai pihak. Oleh karena itu, disampaikan terima kasih dan apresiasi yang sebesar-besarnya kepada:

a. Nassim Dehouche, Mahidol University International College, yang telah membagikan datasetnya di UCI Repository sehingga dapat digunakan dalam penelitian ini.
b. Ibu Dwi Rolliawati, MT yang telah menjadi pembimbing dari suksesnya penelitian ini dari awal hingga akhir, baik dari segi pelaksanaan maupun penulisan laporan.
c. Farah Dwi Wahyuningtyas, Abdillah Arafat, dan Agus Stiawan yang telah melakukan effort dan kerja sama dengan semaksimal mungkin.

6. Daftar Pustaka

[1] Asosiasi Penyelenggara Jasa Internet Indonesia, "Buletin APJII," 2020, vol. 74.

[2] A. Saputra, "Survei Penggunaan Media Sosial di Kalangan Mahasiswa Kota Padang Menggunakan

Teori Uses and Gratifications," *J. Dok. DAN Inf.*, vol. 40, no. 2, p. 207, May 2019, doi: 10.14203/j.baca.v40i2.476.
[3] A. P. Mirsah, "Efektivitas Pemanfaatan Jejaring Sosial (Facebook) Sebagai Media Bisnis Online Dalam Meningkatkan Volume Penjualan (Studi Kasus Makassar Dagang)," p. 107, 2020.
[4] S. Simatupang, E. Efendi, and D. E. Putri, "Facebook Marketplace Serta Pengaruhnya Terhadap Minat Beli," *J. EKBIS*, vol. 22, no. 1, p. 28, Mar. 2021, doi: 10.30736/je.v22i1.695.
[5] Charlie, "Membangun Kepercayaan dan Keterlibatan Konsumen Melalui Live Streaming Social Commerce di Facebook," p. 108, 2020.
[6] D. E. R. Amin and K. Fikriyah, "Pengaruh Live Streaming dan Online Customer Review Terhadap Keputusan Pembelian Produk Fashion Muslim (Studi Kasus Pelanggan TikTok Shop di Surabaya)," vol. 07, no. 01, 2023.
[7] A. Kristianto, "Analisa Performa K-Means dan DBSCAN dalam Clustering Minat Penggunaan Transportasi Umum," *Elkom J. Elektron. Dan Komput.*, vol. 14, no. 2, pp. 368–372, Dec. 2021, doi: 10.51903/elkom.v14i2.551.
[8] R. Adha, N. Nurhaliza, and U. Soleha, "Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia," vol. 18, no. 2, 2021.
[9] M. Putri, C. Dewi, E. P. Siam, G. A. Wijayanti, N. Aulia, and R. Nooraeni, "Comparison of DBSCAN and K-Means Clustering for Grouping the Village Status in Central Java 2020," *J. Mat. Stat. Dan Komputasi*, vol. 17, no. 3, May 2021, doi: 10.20956/j.v17i3.11704.
[10] A. S. Osman, "Data Mining Techniques: Review," *Int. J. Data Sci. Res.*, vol. 2, no. 1, Art. no. 1, Jul. 2019.
[11] Edy Irawan, "Clustering," *School of Computer Science*. https://socs.binus.ac.id/2017/03/09/clustering/ (accessed Nov. 25, 2022).
[12] R. R. Aryanto, A. R. Pratama, and Lizda Iswari, "Studi Komparasi Model Klasifikasi Berbasis Pembelajaran Mesin untuk Sistem Rekomendasi Program Studi," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 5, no. 5, pp. 853–862, Oct. 2021, doi: 10.29207/resti.v5i5.3392.
[13] S. A. H. Bukhari, "What is Comparative Study." Rochester, NY, Nov. 20, 2011. doi: 10.2139/ssrn.1962328.
[14] "UCI Machine Learning Repository: Facebook Live Sellers in Thailand Data Set." https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand (accessed Dec. 10, 2022).
[15] P. S. Bravin, "A review on Data Preprocessing Techniques in Data Mining," vol. 4, no. 5, 2022.
[16] P. W. Setyaningsih and A. Witanti, "Text Mining Repository Untuk Tren Tema Skripsi 2017-2020," *J. INTEK*, vol. 5, no. 2, Feb. 2022, doi: 10.37729/intek.v5i2.2144.
[17] G. J. Kamani, R. S. Parmar, and Y. R. Ghodasara, "Data Normalization in Data Mining Using Graphical User Interface," vol. 30, no. 2, 2019.
[18] "6.3. Preprocessing data," *scikit-learn*. https://scikit-learn/stable/modules/preprocessing.html (accessed Feb. 01, 2023).
[19] A. M. Retta, A. Isroqmi, and T. D. Nopriyanti, "Pengaruh Penerapan Algiritma Terhadap Pembelajaran Pemrograman Komputer," *Indiktika J. Inov. Pendidik. Mat.*, vol. 2, no. 2, p. 126, May 2020, doi: 10.31851/indiktika.v2i2.4125.
[20] H. Februariyanti, J. S. Wibowo, D. B. Santoso, and M. Sukur, "Analisis Kecenderungan Informasi Menggunakan Algoritma Hierarchical Agglomerative Clustering," *INFORMATIKA*, vol. 13, no. 1, p. 9, Jun. 2021, doi: 10.36723/juri.v13i1.247.
[21] W. Widyawati, W. L. Y. Saptomo, and Y. R. W. Utami, "Penerapan Agglomerative Hierarchical Clustering Untuk Segmentasi Pelanggan," *J. Ilm. SINUS*, vol. 18, no. 1, p. 75, Jan. 2020, doi: 10.30646/sinus.v18i1.448.
[22] K. S. Dorman and R. Maitra, "An Efficient K-Modes Algorithm for Clustering Categorical Datasets," *Wiley*, vol. 15, no. 1, pp. 83–97, Sep. 2021, doi: 10.1002/sam.11546.
[23] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
[24] K. Laskhmaiah, S. M. Krishna, and B. E. Reddy, "An Optimized K-means with Density and Distance-Based Clustering Algorithm for Multidimensional Spatial Databases," *Int. J. Comput. Netw. Inf. Secur.*, vol. 13, no. 6, pp. 70–82, Dec. 2021, doi: 10.5815/ijcnis.2021.06.06.
[25] A. K. Fauziyyah, "Analisis Sentimen Pandemi COVID19 Pada Streaming Twitter Dengan Text Mining Python," *J. Ilm. SINUS*, vol. 18, no. 2, p. 31, Jul. 2020, doi: 10.30646/sinus.v18i2.491.
[26] R. J. G. B. Campello, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 2, Mar. 2020, doi: 10.1002/widm.1343.