

Perbandingan Hasil Analisis Clustering Metode K-Means, DBSCAN dan Hierarchical pada Data Marketplace Electronic Phone

¹**Ifatus Sufairoh**, ²**Aulia Cahya Rani**, ³**Khofifah Amalia**, ⁴**Dwi Rolliawati**
Sistem Informasi, Sains dan Teknologi, UIN Sunan Ampel Surabaya
e-mail: ¹h76219022@student.uinsby.ac.id, ²h76219020@student.uinsby.ac.id,
³h06219006@student.uinsby.ac.id, ⁴dwi_roll@uinsby.ac.id
*Penulis Korespondensi

Diterima: 04 April 2023; Direvisi: 12 Juni 2023; Disetujui: 12 Juni 2023

Abstrak

Persaingan yang ketat terjadi di bidang penjualan, salah satunya pada *electronic marketplace* (*e-marketplace*). Pada setiap *e-marketplace* tentu memiliki data penjualan sehari-hari yang semakin lama, data akan bertambah. Bukan hanya digunakan untuk dokumen perusahaan, tetapi juga bisa digunakan serta dikelola sehingga menghasilkan suatu informasi yang bermanfaat dalam meningkatkan promosi dan penjualan produk jika diolah dengan baik. Tujuan dari penelitian ini untuk mengolah data *e-marketplace* dengan cara clustering dan membandingkan hasil yang diperoleh dari setiap model clustering yang dipakai. Model clustering tersebut adalah *K-Means*, *DBSCAN*, dan *Hierarchical*. Hasil yang didapatkan yaitu dari ketiga metode clustering yang digunakan didapat model terbaik yaitu *Hierarchical* dengan jumlah cluster 2 dari silhouette sebesar 0.944473. Dari hasil tersebut dapat diketahui pengelompokan penjualan handphone dengan merk terbanyak dan hasil clustering dapat menjadi perbandingan metode yang paling optimal.

Kata kunci: marketplace, clustering, k-means, dbscan, hierarchical

Abstract

Intense competition occurs in the field of sales, one of which is in *electronic marketplaces* (*e-marketplaces*). Every *e-marketplace* certainly has daily sales data which is getting longer, the data will increase. It is not only used for company documents, but can also be used and managed so as to produce useful information in increasing product promotion and sales if it is processed properly. The purpose of this study is to process *e-marketplace* data by means of clustering and compare the results obtained from each clustering model used. The clustering models are *K-Means*, *DBSCAN*, and *Hierarchical*. The results obtained are from the three clustering methods used, the best model is *Hierarchical* with the number of clusters 2 of the silhouette of 0.944473. From these results it can be seen that the grouping of cell phone sales with the most brands and the clustering results can be the most optimal comparison method.

Keywords: electronic marketplace, clustering, k-means, dbscan, hierarchical

1. PENDAHULUAN

Di bidang penjualan, persaingan yang ketat di dunia usaha menuntut pengembang untuk mengembangkan strategi yang dapat meningkatkan penjualan serta pemasaran pada produk penjualan. Seperti penggunaan data penjualan sehari-hari. Data tersebut terus bertambah seiring berjalannya waktu. Selain berperan sebagai arsip perusahaan, data ini memiliki potensi untuk

digunakan dan dimanfaatkan sebagai informasi berharga dalam meningkatkan penjualan dan promosi produk [1]. Salah satu contohnya adalah berbagai jenis barang elektronik, seperti Handphone, yang hadir dengan beragam fungsi untuk memudahkan manusia dalam berbagai aktivitas. Handphone adalah barang yang sangat dibutuhkan karena dapat membantu dalam aktivitas manusia [2].

Handphone merupakan alat komunikasi elektronik yang dapat dibawa dengan mudah serta mempunyai kemampuan untuk mengirimkan pesan berbentuk suara [3]. Dalam kehidupan sehari-hari, kita tidak bisa lepas dari perangkat telepon genggam (ponsel), terlebih dengan terus meningkatnya handphone sehingga mempunyai bermacam-macam manfaat. Tidak sekedar alat komunikasi saja tetapi menjadi perlengkapan dengan kegunaan yang lain seperti wadah berita, hiburan, berbisnis, dan lainnya. Saat ini sebutannya smartphone ataupun ponsel pintar. Istilah untuk handphone yang dapat digunakan untuk melaksanakan banyak perihal [4].

Ketika handphone belum mempunyai banyak kegunaan, handphone sudah melewati proses yang panjang sejak awal kemunculannya. Peningkatan dan perkembangan teknologi semakin meningkat begitu juga merek handphone yang diperjualbelikan di Indonesia. Sebagian ada yang dari berbagai negara dan sebagian lagi dari produk Indonesia. Harga yang diperjualkan pastinya berbeda antara satu dengan yang lainnya. Di antara sekian banyak ponsel yang dijual, tidak ada kekurangan ponsel yang sangat dicari karena kualitasnya yang tinggi dan daya tahan pemakaiannya. Ada banyak merek ponsel yang disukai masyarakat Indonesia, mulai dari remaja hingga manula.

Pada penelitian ini, data mining diterapkan pada data penjualan handphone pada marketplace menggunakan metode *clustering*. Data mining adalah merupakan suatu proses yang digunakan untuk mengidentifikasi pola dan pengetahuan yang menarik dari sejumlah besar data [5]. Selain itu menggambarkan sekumpulan teknik yang ditujukan untuk mengungkap pola yang tidak diketahui dalam data yang dikumpulkan [6]. Beberapa penelitian telah melakukan klusterisasi pada data penjualan handphone pada marketplace. Tujuan dilakukannya penelitian ini adalah untuk segmentasi pasar penjualan handphone dari kalangan remaja hingga orang tua.

Algoritma K-Means, DBSCAN, dan algoritma Hierarchical termasuk dalam kategori pengelompokan yang tidak diawasi (*Unsupervised Learning*) dan digunakan untuk mengelompokkan data menjadi beberapa kelompok [7]. Algoritma K-Means melakukan partisi data ke dalam cluster yang nantinya data dengan karakteristik sama akan digabungkan pada cluster yang sama begitu sebaliknya untuk karakteristik yang berbeda [8]. DBSCAN melakukan cluster pada data yang berdekatan dan data yang berjauhan tidak menjadi anggota cluster [9]. Algoritma Hierarchical menggabungkan dua atau lebih objek berdasarkan nilai kemiripan yang paling dekat. Kemudian hasil penggabungan tadi digabungkan lagi dengan obyek lain berdasarkan nilai kemiripan kedua dan seterusnya hingga membentuk konstruksi hirarki [10]. Penerapan tiga algoritma ini guna mengelompokkan penjualan handphone dengan merk terbanyak dengan data yang diperoleh dari website dataset Kaggle. Hasil clustering ini dapat menjadi perbandingan metode yang paling optimal. Berdasarkan permasalahan yang dijelaskan, penulis merasa tertarik untuk menyelidiki topik ini dalam sebuah penelitian yang berjudul "Perbandingan Analisis Pengelompokan Metode K-Means, DBSCAN, dan Hierarchical pada Data Marketplace Electronic Phone". Penelitian ini merupakan pengembangan dari penelitian sebelumnya dengan penggunaan data dan atribut yang lebih luas. Jumlah data yang digunakan dalam penelitian ini adalah sebanyak 2480 entri.

2. METODE PENELITIAN

2.1. Data

Dalam penelitian ini, data yang digunakan adalah Marketplace Electronic (Phone) yang diambil dari website dataset Kaggle dengan jumlah data sebanyak 26015. Data yang digunakan

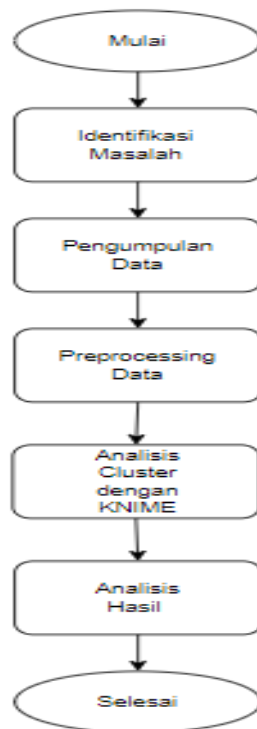
terdapat 16 atribut yaitu : date, sku_name, product_id, merchant_code, merchant_badge, product_type, merchant_city, cod, price, stock, daily_gmv, daily_sold, daily_view, rating, total_review, dan total_sold. Berikut ini akan ditampilkan beberapa data yang digunakan dalam penelitian ini.

Tabel 1. Data Marketplace Electronic (Phone)

Date	...	Product_id	...	Product_type	...	Price
3/11/2020	E1	Huawei nova 5T	4500000
3/26/2020	F1	Huawei nova 5T	4999000
3/23/2020	F2	Huawei nova 5T	4589900
4/29/2020	E2	Huawei nova 5T	4698500

2.2. Pengolahan Data

Setelah memperoleh data, data tersebut kemudian diproses menggunakan *tools* KNIME untuk melakukan proses clustering. Proses clustering yang digunakan pada penelitian ini antara lain algoritma K-Means, DBSCAN, dan Hierarchical. Berikut merupakan langkah-langkah menyelesaikan penelitian ini ada pada gambar di bawah ini.
















Gambar 1. Langkah-langkah Penelitian

Ada beberapa tahapan dalam penelitian ini yaitu dimulai dari identifikasi masalah yang ada, selanjutnya pengumpulan data, lalu *preprocessing* data digunakan untuk menghilangkan gangguan (*noise*), nilai yang hilang (*missing value*), dan juga data yang tidak konsisten, kemudian pengolahan data dengan algoritma K-Means, DBSCAN, dan Hierarchical, *Clustering* di aplikasi KNIME, dievaluasi kualitas clusteringnya, dan disimpulkan hasilnya. Terdapat beberapa node yang dibutuhkan dalam melakukan clustering dengan algoritma K-Means, DBSCAN, dan Hierarchical menggunakan *tools* KNIME, antara lain *CSV Reader*, *Missing Value*, *Category to Number*, *Partitioning*, *k-Means*, *Color Manager*, *Shape Manager*, *Scatter Plot*, *Silhouette Coefficient*, *Cluster Assigner*, *Numeric Distances*, *DBSCAN*, *Category to Number*, *Distance*

Matrix Calculate, Hierarchical Clustering (DistMatrix), Hierarchical Cluster View, Hierarchical Clustering, dan Silhouette Coefficient seperti gambar tabel 2 dibawah ini.

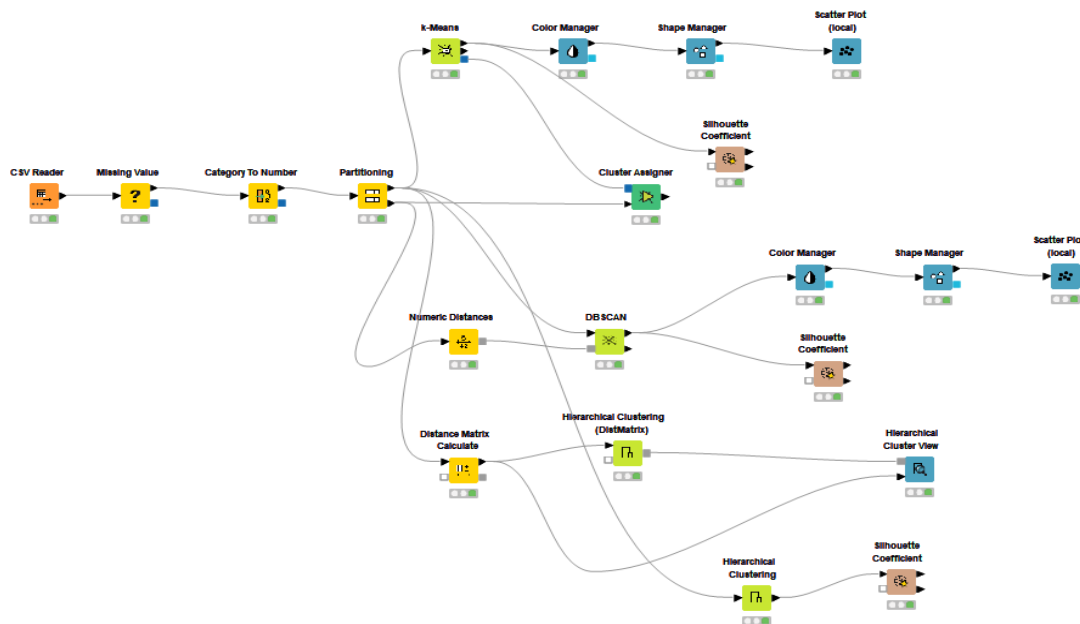
Tabel 2. Node KNIME

No	Node	Nama Node	Deskripsi
1		CSV Reader	Membaca file dengan format CSV.
2		Missing value	Untuk membantu menangani nilai yang hilang yang ditemukan di sel tabel input.
3		Category to number	Untuk mengambil kolom dengan data nominal dan memetakan setiap kategori ke bilangan bulat.
4		Partitioning	Membagi tabel menjadi dua partisi.
5		K-means	Untuk menampilkan pusat cluster untuk sejumlah cluster yang telah ditentukan.
6		Color manager	Untuk memberikan warna dengan nominal (nilai yang mungkin harus tersedia) atau kolom numerik (dengan batas bawah dan atas).
7		Shape manager	Untuk menampilkan bentuk (berbeda) dalam setiap nilai atribut dari satu kolom nominal.
8		Scatter plot	Tipe grafik untuk menggambarkan data dengan menggunakan koordinat cartesian.
9		Silhouette coefficient	Untuk menghitung Koefisien Silhouette hasil clustering.
10		Cluster assigner	Untuk menugaskan data baru ke kumpulan prototipe yang ada.
11		Numeric Distances	Untuk menghitung Koefisien Silhouette untuk hasil pengelompokan yang disediakan.
12		DBSCAN	Untuk melakukan clustering dataset menggunakan DBSCAN.
13		Distance Matrix Calculate	Menghitung nilai jarak untuk semua pasangan baris dari tabel input. Hasilnya ditambahkan ke tabel input sebagai satu kolom yang berisi nilai vektor jarak.
14		Hierarchical Clustering (DistMatrix)	Untuk mengelompokkan data input menggunakan matriks jarak.
15		Hierarchical Cluster View	Mengambil hierarki pohon klaster dan tabel masukan yang sama, yang telah digunakan untuk membuat pengelompokan dan memvisualisasikan dendrogram klaster dan plot jarak pada semua level yang dibuat.

No	Node	Nama Node	Deskripsi
16		Hierarchical Clustering	Untuk mengelompokkan data.

3. HASIL DAN PEMBAHASAN

Setelah data diproses menggunakan *tools* KNIME, kemudian akan ditampilkan hasil clustering dalam bentuk diagram dengan tiga hasil untuk perbandingan yaitu hasil dari metode K-Means, metode DBSCAN, dan metode Hierarchical yang ditampilkan pada gambar 2.



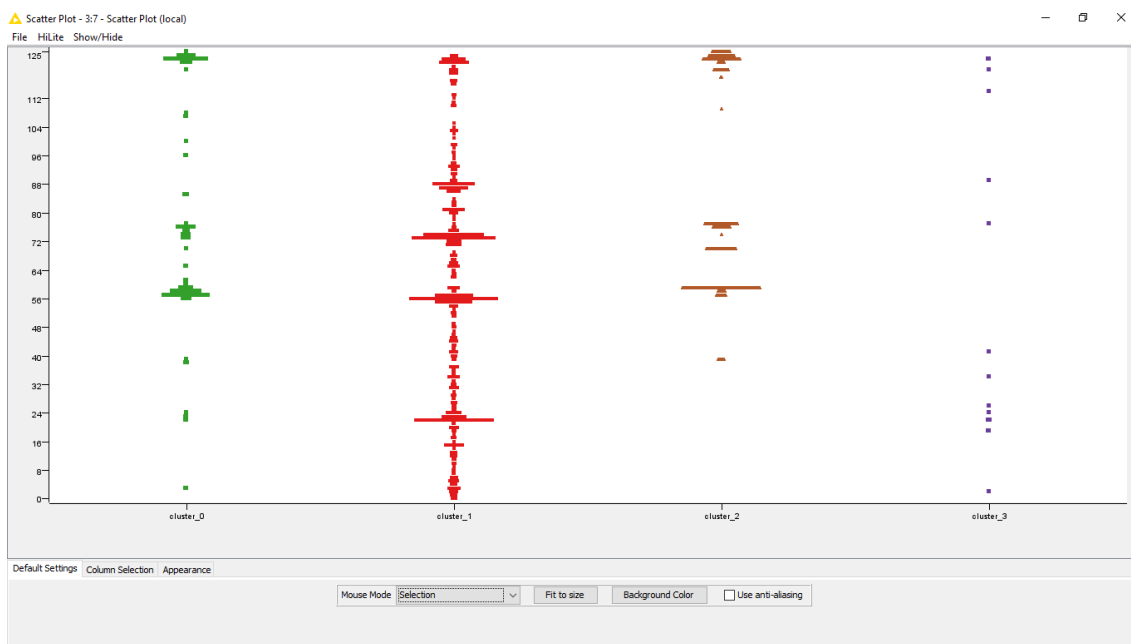
Gambar 2. Workflow

Tahap data preprocessing dimulai dengan menggunakan node CSV reader kemudian missing value, category to number, partitioning. Pada metode clustering K-Means membutuhkan node K-Means untuk menentukan jumlah cluster yang akan dibuat. Pengujian clustering K-Means pada penelitian ini dilakukan sebanyak lima kali dengan jumlah cluster yang berbeda-beda dengan pengujian silhouette coefficient. Pada node silhouette coefficient didapatkan nilai rata-rata keseluruhan silhouette coefficient yang ditunjukkan dalam tabel 3 berikut.

Tabel 3. Hasil Silhouette Coefficient K-Means

Jumlah Cluster	Overall Mean Silhouette Coefficient
2	0.690601
3	0.617352
4	0.631843
5	0.609948
6	0.319187

Pada tabel 3 di atas dipaparkan hasil performa terbaik dengan metode clustering K-Means pada jumlah cluster 2. Dalam melakukan visualisasi data pada metode clustering K-Means digunakan tiga node mulai dari color manager, shape manager, dan scatter plot. Scatter plot hasil clustering K-Means pada sumbu x dan y merupakan jumlah tipe produk pada tiap cluster. Berikut hasil scatter plot clustering K-Means dengan jumlah cluster 2 pada gambar 3.



Gambar 3. Scatter Plot Clustering K-Means Data Phone

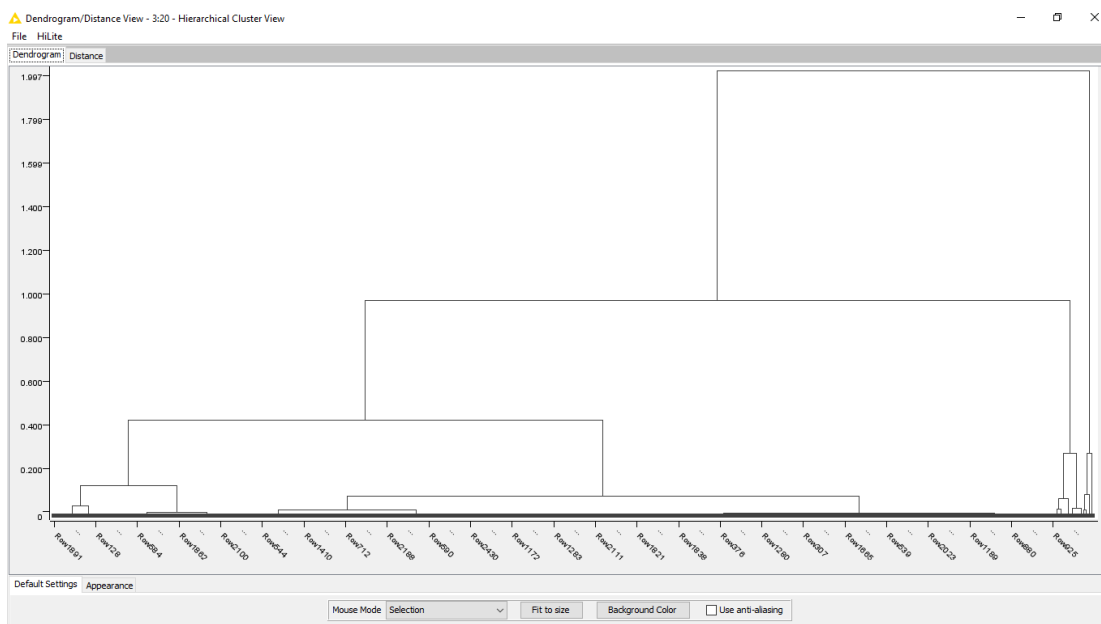
Pemodelan *clustering* yang kedua menggunakan metode DBSCAN pada KNIME yang dimulai dengan *numeric distance* untuk menghitung *koefisien Silhouette*, kemudian node DBSCAN yang digunakan untuk menjalankan metode DBSCAN dan divisualisasikan menggunakan node color manager untuk memberi warna lalu ditampilkan pada node scatter plot. Selanjutnya untuk menghitung *Koefisien Silhouette* hasil *clustering* pada metode DBSCAN digunakan node “*silhouette coefficient*”. *Output* dari node “*silhouette coefficient*” berupa nilai *overall mean silhouette coefficient* yang ada pada tabel 4 dibawah ini.

Tabel 4. Hasil Silhouette Coefficient DBSCAN

Jumlah Cluster	Overall Mean Silhouette Coefficient
2	0.283192
3	0.256690
4	0.239173
5	0.219255
6	0.137512

Pemodelan yang terakhir yaitu metode Hierarchical. Pada metode ini dibagi menjadi dua proses antara lain menampilkan dendrogram dan menentukan jumlah cluster. Dalam menampilkan dendogram diperlukan *node distance matrix calculate*, *hierarchical clustering (distmatrix)* dan *hierarchical cluster view*. Hasil dendrogram pada *clustering hierarchical* ditunjukkan pada gambar 4.

Fase metode *clustering hierarchical* dibagi menjadi dua untuk menampilkan dendrogram dan menentukan jumlah cluster. Dalam menampilkan dendogram diperlukan *node distance matrix calculate*, *hierarchical clustering (distmatrix)* dan *hierarchical cluster view*. Dendrogram hierarchical clustering pada gambar 3 menunjukkan anggota cluster dan jarak kombinasi cluster. Anggota cluster berada pada sumbu x dan jarak kombinasi cluster diskalakan pada sumbu y. Untuk mengetahui anggota cluster mana yang akan terbentuk ditentukan dengan jumlah *clade* terhitung dari atas dendogram. Misal apabila dibentuk 2 cluster, maka cluster 1 adalah turunan dari *clade* terdekat sumbu 0 dan cluster 2 adalah turunan dari *clade* setara yang lain. Untuk menentukan jumlah cluster digunakan node hierarchical clustering.



Gambar 4. Dendrogram Hierarchical Clustering Data Phone

Pengujian hierarchical clustering dilakukan sebanyak lima kali dengan menentukan jumlah cluster mulai dari dua hingga enam cluster. Untuk mendapatkan nilai performa dari setiap cluster digunakan node silhouette coefficient, didapatkan hasil pada tabel 3 berikut dimana nilai rata-rata keseluruhan silhouette coefficient terbaik dengan 2 cluster.

Tabel 5. Hasil Silhouette Coefficient Hierarchical

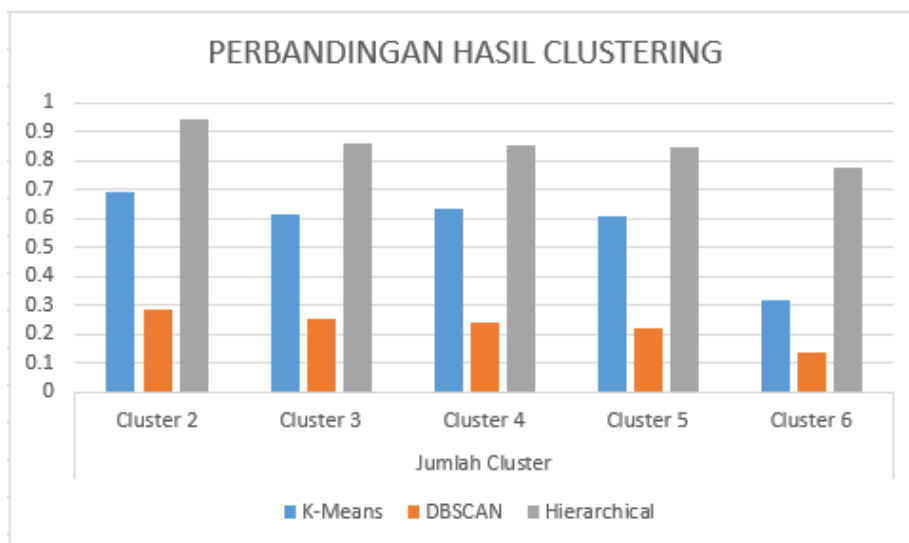
Jumlah Cluster	Overall Mean Silhouette Coefficient
2	0.944473
3	0.857908
4	0.855926
5	0.850590
6	0.778426

Perbandingan algoritma K-Means, DBSCAN dan Hierarchical dilakukan dengan menentukan jumlah cluster yang paling optimal dengan cara dilihat dari nilai Silhouette Coefficient. Hasil clustering menggunakan data kasus marketplace Phone dengan metode K-Means, DBSCAN, dan Hierarchical menghasilkan 6 cluster dengan nilai yang berbeda-beda. Perbandingan nilai Silhouette coefficient tertinggi dapat dilihat pada tabel berikut :

Tabel 6. Perbandingan Hasil

Metode Clustering	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
K-Means	0.690601	0.617352	0.631843	0.609948	0.319187
DBSCAN	0.283192	0.256690	0.239173	0.219255	0.137512
Hierarchical	0.944473	0.857908	0.855926	0.850590	0.778426

Tabel 6 diatas merupakan perbandingan hasil pengujian cluster terhadap algoritma K-Means, DBSCAN dan Hierarchical menggunakan nilai Silhouette Coefficient. Maka pada penelitian ini, algoritma Hierarchical memiliki nilai cluster lebih baik dibandingkan dengan algoritma lain yang ditunjukkan pada gambar dibawah 5.



Gambar 5. Perbandingan Hasil Clustering

4. KESIMPULAN

Kesimpulan yang didapatkan adalah berdasarkan hasil analisis perbandingan dengan metode K-Means, DBSCAN dan Hierarchical dalam clustering data marketplace electronic phone didapat model terbaik yaitu Hierarchical dengan jumlah cluster 2 dari silhouette sebesar 0.944473. Dengan adanya informasi ini, dapat diketahui pengelompokan penjualan handphone dengan merk terbanyak dan hasil clustering dapat menjadi perbandingan metode yang paling optimal.

5. SARAN

Untuk penelitian berikutnya, penting untuk mengatur Algoritma Komputasi yang berbeda. Selain itu, penambahan dataset dan pusat kluster juga dapat dilakukan untuk menghasilkan lebih banyak kluster.

DAFTAR PUSTAKA

- [1] U. D. Soer and B. K. L. Batu, "Implementasi Algoritma Apriori Untuk Prediksi Penjualan Produk Packaging Di PT. Printec Perkasa II," vol. 9, no. 3, 2019.
- [2] F. P. A. Hasibuan, S. Sumarno, and I. Parlina, "Penerapan K-Means pada Pengelompokan Penjualan Produk Smartphone," *SATESI J. Sains Teknol. Dan Sist. Inf.*, vol. 1, no. 1, pp. 15–20, Sep. 2021, doi: 10.54259/satesi.v1i1.3.
- [3] L. A. Suryanita and P. K. Arieska, "Analisis Cluster Persepsi Konsumen Terhadap Produk Handphone Untuk Menentukan Strategi Promosi Penjualan Di ZICOMS," *J Stat. J. Ilm. Teori Dan Apl. Stat.*, vol. 8, no. 1, Dec. 2015, doi: 10.36456/jstat.vol8.no1.a309.
- [4] I. Nuryani and D. Darwis, "Analisis Clustering Pada Pengguna Brand HP Menggunakan Metode K-MEANS," vol. 1, no. 1, 2021.
- [5] P. A. Rahayuningsih and R. Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner," vol. 6, no. 1, 2018.
- [6] M. A. Senubekti and L. A. Puspita Dewi, "Prinsip Klasifikasi Dan Data Mining Dengan Algoritma C4.5," *NUANSA Inform.*, vol. 16, no. 2, pp. 87–93, Jul. 2022, doi:

- 10.25134/nuansa.v16i2.5834.
- [7] A. A. Herlambang, M. A. Murti, and C. Setianingsih, "Pengelompokan Data Penggunaan Energi Listrik Menggunakan Algoritma Mini Batch K-Meansclustering," vol. 9, no. 5, 2022.
- [8] P. Silitonga and I. S. Morina, "Klusterisasi Pola Penyebaran Penyakit Pasien Berdasarkan Usia Pasien Dengan Menggunakan K-Means Clustering," *J. Times*, vol. 6, no. 2, 2017.
- [9] D. K. Alfiki Astutik, A. Indrasetianingsih, and F. Fitriani, "Penerapan Text Mining pada Analisis Sentimen Pengguna Twitter Layanan Transportasi Online Menggunakan Metode Density Based Spatial Clustering of Applications With Noise (DBSCAN) dan K-Means," *J Stat. J. Ilm. Teori Dan Apl. Stat.*, vol. 15, no. 1, Jul. 2022, doi: 10.36456/jstat.vol15.no1.a5983.
- [10] T. Alfina, B. Santosa, and J. A. R. Hakim, "Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)," vol. 1, p. 5, 2012.
-