JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

PERBANDINGAN HASIL METODE CLUSTERING K-MEANS, DB SCANNER & HIERARCHICAL UNTUK ANALISA SEGMENTASI PASAR

Syafrina Dyah Kusuma Wardani[1], **Amalia Salsabilla Ariyanto**[2], **Masfufahtul Umroh**[3], dan **Dwi Rolliawati**[4]
[1,2,3,4]Program Studi Sistem Informasi, UIN Sunan Ampel Surabaya, Indonesia
Email: h76219033@student.uinsby.ac.id[1], h96219040@student.uinsby.ac.id[2],
h96219050@student.uinsby.ac.id[3], dwi_roll@uinsby.ac.id[4]

Abstrak

Segmentasi pasar merupakan strategi pengelompokan calon konsumen berdasarkan persepsi yang sama antara kebutuhan dan keinginan. Dalam strategi pemasaran, segmentasi pasar sangat wajib untuk diterapkan karena penentuan segmentasi pasar merupakan dasar dari adanya pemasaran. Namun seringkali terdapat hambatan dalam melakukan segmentasi pasar seperti tidak ada pembaruan segmentasi, menghiraukan calon konsumen, tidak mempunyai banyak data. Sehingga tujuan penelitian ini untuk menentukan segmentasi pasar dilakukan dengan clustering yakni mengelompokkan data sesuai karakteristik dari konsumen dan membandingkan hasil clustering model terbaik. Data yang digunakan merupakan data segmentasi pasar penjualan otomotif yang diambil dari sumber kaggle.com. Peneltian ini menggunakan metode clustering K-Means, DBSCAN, serta Hierarchical, dimana pada proses pemodelannya menggunakan software KNIME. Adapun hasil dari penelitian ini yakni metode K-Means adalah metode yang baik untuk penerapan clustering. Hal ini didapatkan dari nilai Mean Silhouette Coefficient K-Means mendekati 1 yakni 0.716, sedangkan DBSCAN 0.296 dan Hierarchical 0.301
Kata Kunci: *Data Mining, K-Means, DBSCAN, Hierarchical, Silhouette Coefficient.*

Abstract

Market segmentation is a strategy for grouping potential customers based on the same perception of needs and wants. In the marketing strategy, market segmentation is mandatory to implement because the determination of market segmentation is the basis of marketing. However, there are often obstacles in conducting market segmentation such as no segmentation updates, ignoring potential customers, and not having a lot of data. So the purpose of this research is to determine market segmentation by clustering, namely grouping data according to the characteristics of consumers and comparing the results of the best clustering models. The data used in automotive sales market segmentation data is taken from kaggle.com. This research uses the K-Means, DBSCAN, and Hierarchical clustering methods, where the modelling process uses KNIME software. The results of this study are that the K-Means method is a good method for implementing clustering. This is obtained from the K-Means Mean Silhouette Coefficient close to 1, namely 0.716, while DBSCAN is 0.296 and Hierarchical is 0.301
KeyWords: *Data Mining, K-Means, DBSCAN, Hierarchical, Silhouette Coefficient.*

I. PENDAHULUAN

Seiring berkembangnya data dan informasi memungkinkan suatu data atau informasi mudah diakses dan didapatkan oleh siapapun. Adanya data yang besar dan keterbukaan informasi mensupport manusia dalam menggunakan data struktur atau tidak terstruktur. Ilmu yang membahas mengenai data serta pengolahan data merupakan ilmu *Data Mining*. *Data mining* biasa digunakan dalam memastikan bentuk dari suatu data yang dapat digunakan dalam pengambilan keputusan pada suatu perusahaan atau organisasi [1].

Suatu data atau informasi sangat dibutuhkan di berbagai sektor salah satunya ekonomi dan bisnis. Berhasil tidaknya usaha bisnis dapat ditentukan dari beberapa sudut pandang termasuk bagaimana bisnis itu dapat mempengaruhi konsumen yang dituju. Berdasarkan hal itu maka diperlukannya pemahaman mengenai segmentasi pasar. Bahwasannya segmentasi harus dilakukan sebelum menentukan target pasar. Pengelompokkan data segmentasi pasar merupakan salah satu solusi untuk mengetahui karakteristik target pasar [2].

Pemasaran merupakan pekerjaan yang ruang lingkupnya adalah memasarkan berbagai produk. Tugas pemasaran dalam suatu perusahaan sangat berpengaruh dalam berhasil tidaknya kerja perusahaan tersebut. Pemasaran memegang peranan yang kuat dalam memasarkan produk suatu perusahaan agar perusahaan dapat bertahan. Sehingga hanya dengan pemasaran yang andal sebuah perusahaan dapat memperoleh pijakan yang kuat [3].

Segmentasi pasar merupakan strategi pengelompokan calon konsumen berdasarkan persepsi yang sama antara kebutuhan dan keinginan. Dalam strategi pemasaran, segmentasi pasar sangat wajib diterapkan karena penentuan segmentasi pasar dasar dari adanya pemasaran. Segmen pasar juga didefinisikan sebagai pengelompokan pembeli yang memiliki perbedaaan pada kebutuhan, karakteristik, atau perilaku yang berbeda, dimana kemungkinan membutuhkan produk [4]. Segmentasi pasar adalah tahapan memilah pasar menurut kelompok konsumen yang relatif homogen, dan setiap kelompok konsumen dapat digunakan sebagai tujuan yang ingin dicapai perusahaan melalui strategi pemasaran [5].

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

Dengan mengamati kecenderungan pembelian konsumen untuk menentukan pola penjualan, jika dianalisis dan ditangani secara baik, dapat memudahkan untuk mengetahui produk mana yang laris manis serta mana yang tidak laris, jadi dapat dilakukan pengendalian persediaan. Ini dapat digunakan sebagai masukan untuk strategi pemasaran perusahaan [6].

Tujuan penelitian ini yakni untuk mengetahui kinerja dari metode K-Means, DB SCANNER, dan Hierarchical dengan cara mengelompokkan data segmentasi pasar produk otomotif. Hasil dari pengelompokkan ditujukan untuk mengetahui produk otomotif yang paling laku terjual sesuai dengan pengelompokan yang ada, kemudian akan dilakukan perbandingan hasil dari masing-masing metode yang terbaik.

Clustering adalah metode analisis data, biasanya salah satu teknik *data mining*, yang membagi data menjadi suatu wilayah dengan menggunakan kesamaan karakteristik suatu wilayah, sehingga data dan karakteristik wilayah tersebut berbeda. *Clustering* secara umum mengacu pada membagi sekumpulan data menjadi kelompok-kelompok sedemikian rupa akibatnya objek pada suatu kelompok mempunyai kesamaan serta perbedaan yang banyak dari kelompok objek lainnya [7]. Pengelompokan juga disebut Klasifikasi Tanpa Pengawasan (*Unsupervised Classification*), karena pengelompokan membutuhkan lebih banyak pembelajaran dan perhatian. Persamaan dan perbedaan dalam *clustering* biasanya didasarkan pada nilai atribut objek, dan bisa juga dalam bentuk perhitungan jarak [8]. Analisis *cluster* suatu tahapan memilah kumpulan objek data nebjadi subset, dimana subset merupakan klaster sedemikian rupa yang mana objek pada klaster sama dengan yang lain namun berbeda pada objek dengan klaster yang lain [9].

K-Means adalah algoritma yang diimplementasikan dalam mengelompokkan dengan maksud menyederhanakan dan efisiensi. K-Means adalah algoritma *clustering* atau pengelompokan yang mendasari sifat metode *non-hierarchy* yang memisahkan data serta mengelola beberapa kelompok yang memiliki kemiripan nilai atau nilai yang sama [10]. K-Means merupakan metode pengelompokan data *non-hierarkis* (batas) yang mengelompokkan data menjadi dua kelompok ataupun lebih dari dua kelompok. K-Means ini membagi data pada kelompok, dan karakter data yang sama akan diklasifikasikan pada kelompok yang sama, begitupun sebaliknya karakter data yang berbeda akan diklasifikasikan pada kelompok lain [11]. Pengelompokan memiliki tujuan yakni meminimalisir fungsi tujuan yang ditetapkan dalam tahapan pengelompokan. Secara umum tipe pada kelompok akan diminimalkan dan tipe antar kelompok akan dimaksimalkan [12].

Langkah algoritma K-Means menurut Sarwono yakni dengan menetapkan jumlah *cluster*, menyebarkan data *cluster* secara *random* atau acak dengan menggunakan rumus yang ada, lalu mengelompokkan data ke *cluster* melalui jarak yang pendek, selanjutnya menghitung pusat yang ada pada *cluster* dengan menggunakan rumus yang ada dan selanjutnya mengulangi langkah - langkah hingga tidak terdapt data berpindah ke *cluster* lainnya [13].

DBSCAN atau *Density Based Spatial Clustering of Applications with Noise* merupakan algoritma yang mendapatkan *core* sample dengan memperlebar *cluster* pada sampel yang ada [14]. Algoritma tersebut memiliki dua parameter utama untuk menentukan *cluster*. Skala atau patokan pertama adalah menetapkan jumlah minimum titik yang bisa dikategorikan dengan *core sample*. Skala atau patokan ini menentukan tingkatan margin kebisingan. Skala atau patokan kedua adalah ε.

Hierarchical merupakan sebuah algoritma pengelompokan data yang diawali dengan setiap observasi sebagai klasternya sendiri dan dilanjutkan dengan mengelompokkan. observasi ke dalam kelompok yang lebih besar. Pengelompokan *hierarchical* seperti itu disebut dengan pendekatan *bottom-up* [15]. Metode analisis *cluster* secara *hierarchical* dapat dibedakan menjadi dua yaitu *Devise*, dimana dimulai dengan membuat satu klaster yang memuat seluruh objek atau pengamatan, kemudian objek yang paling berbeda 'kemiripannya' dipisahkan dari klaster demikian seterusnya hingga setiap objek terpisah dan menjadi satu klaster [16]. *Agglomerative*, yaitu setiap objek dianggap sebagai satu klaster lalu dua objek yang memiliki 'kemiripan' disatukan menjadi satu klaster demikian seterusnya hingga terbentuk satu klaster yang terdiri atas keseluruhan objek [17].

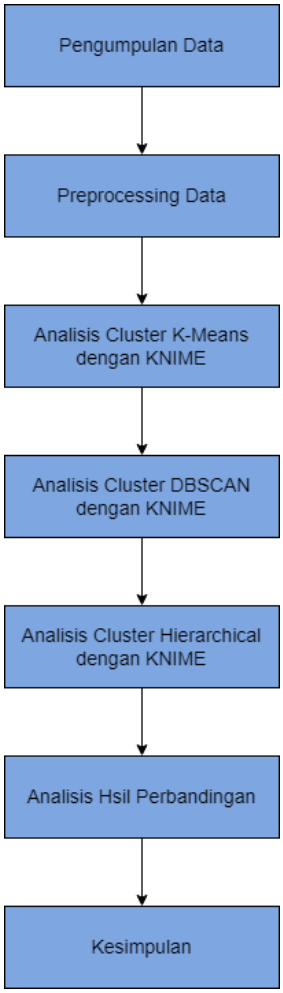
Silhouette Coefficient adalah metode untuk mengevaluasi hasil pengelompokan yang menggabungkan metode kohessian dengan metode pemisahan. Dimana kohessian mengukur jumlah keseluruhan objek yang ada pada *cluster* sedangkan separation mengukur jarak rata-rata di setiap objek pada *cluster*. Masing-masing *cluster* dihitung dengan menggunakan *silhouette* [18]. Nilai dari *silhouette* didefinisikan dengan *sil* (k) dihitung dengan menggunakan rata-rata dari *silhouette* pada *cluster*. Nilai dari hasil *Silhouette Coefficient* berkisar -1 hingga dengan 1, dimana yang mengarah nilai 1 termasuk pengelompokan data klaster terbaik. Sedangkan nilai yang mengarah pada nilai -1 termasuk pengelompokan data klaster yang tidak baik atau buruk [19].

KNIME merupakan *software* yang bersifat *open source* atau perangkat lunak sumber terbuka yang mudah diakses. KNIME banyak digunakan sebagai *platform* analisis data, pelaporan, dan integrasi yang dapat menggabungkan berbagai elemen pembelajaran mesin dan penambangan data dengan konsep diagram data. KNIME sendiri memiliki banyak keunggulan pada fitur-fiturnya karena dengan mudah dapat memodelkan setiap langkah *machine learning* dengan banyak bahasa, seperti python, SQL, CRUDA, C, R. PyTorch, dan yang lainnya. KNIME membuat pemahaman data dan mengembangkan alur kerja ilmu data dan komponen yang dapat digunakan kembali dapat diakses oleh menjadi intuitif, transparan, dan terus menggabungkan teknologi baru [20].

II. METODE

Data yang diaplikasikan dalam penelitian ini yakni data sekunder. Data set yang diterapkan pada penelitian ini bersumber dari Kaggle yakni https://www.kaggle.com/datasets/ankitchahal1/sales-data. Penelitian ini menggunakan *software* KNIME untuk pengolahan data dilakukan dengan menerapkan metode K-Means, DBSCAN, dan Hierarchical untuk melakukan analisis data mengenai segmentasi pasar. Adapun langkah-langkah menyelesaikan penelitian. ini ada pada Gambar 1 di bawah ini.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796



Gambar 1: Tahapan Penelitian

Berikut adalah deskripsi singkat terkait alur tahapan penelitian pada Gambar 1 diatas :

1) Pengumpulan Data
   Tahapan proses pengumpulan data, dimana pada penelitian ini mencari sumber data melalui situs website kaggle.com. Data yang didapatkan sifatnya sekunder mengenai segmentasi pasar penjualan produk otomotif yang berjumlah sebanyak 2000 data dengan 25 Atribut.
2) Preprocessing Data
   Tahapan preprocessing data ini adalah *Data Cleaning* dengan menerapkan *Missing Value* dan *Data Transformation* dengan teknik *Normalization*. Tahapan preprocessing data ini adalah proses membersihkan atau merapikan *data set* dengan menggunakan *software* KNIME dengan menerapkan *missing value* dan *Data transformation* dengan teknik *Normalization*, yang bertujuan untuk menghilangkan data yang tidak lengkap atau *noise* agar pada saat proses pemodelan *clustering* data lebih akurat. Pada dataset yang digunakan ini terdapat 3 atribut yang *missing value* yakni atribut *Addressline2*, *State*, dan *Postalcode*.
3) Analisis *Cluster* K-Means dengan KNIME
   Tahapan analisis *cluster* metode algoritma K-Means dengan penerapan *software* KNIME ini diterapkan dengan menggunakan *node-node* yang terkait dengan metode K-Means. Langkah algoritma K-Means menurut Sarwono yakni sebagai. berikut :
   a. Langkah pertama adalah menetapkan jumlah cluster.
   b. Lalu menyebarkan data cluster secara random atau acak dengan menggunakan rumus :

$$Ci = Min + \frac{(i-1) * (max - min)}{n} + \frac{(mox - min)}{2 * n} \quad (1)$$

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

Keterangan :
Ci : *Centroid* pada kelas i
Min : Nilai min kelas kontinyu
Max : Nilai max kelas diskrit
n : Jumlah kelas diskrit

c. Mengelompokkan data ke *cluster* jarak minimum.
d. Menghitung pusat yang ada pada *cluster*, dengan menggunakan rumus :

$$Be = \frac{\sqrt{(Oi - Ti)^2}}{(Oj - Tj)^2} \tag{2}$$

Keterangan :
Be : Jarak dari data ke pusat
O : Record data
T : Centroid data

e. Ulangi langkah diatas secara berulang sampai dengan data tidak berpindah pada cluster yang lain.

4) Analisis *Cluster* DBSCAN dengan KNIME
Tahapan analisis *cluster* metode algoritma DBSCAN dengan menggunakan *software* KNIME ini diterapkan dengan menggunakan *node-node* yang terkait dengan metode DBSCAN. Berikut ini merupakan langkah-langkah penggunaan algoritma DBSCAN.
a. Tetapkan nilai patokan $MinPts$ dan Eps.
b. Tetapkan nilai p dengan acak atau titik awal.
c. Jumlah Eps terhadap nilai p menggunakan rumus jarak *euclidean* berikut.

$$d_{ij} = \sqrt{\sum_{a}^{p}(x_{ia} - x_{ja})^2} \tag{3}$$

Dimana x_{ia} merupakan variabel ke 1 dari objek i (i=1, …, n; a=1, …, p) dan d_{ij} merupakan nilai *euclidean distance*.
d. *Cluster* terbentuk saat *node* mencukupi nilai Epslebih dari nilai $MinPts$ serta *node* p sebagai *core point*.
e. Ulangi tahapan 3 - 4 sampai proses seluruh *node*. Proses akan dilanjutkan ke titik yang lain apabila nilai p merupakan titik *border*, tidak memiliki titik *density reachable*.

5) Analisis *Cluster Hierarchical* dengan KNIME
Tahapan analisis *cluster* metode algoritma *Hierarchical* dengan menggunakan *software* KNIME ini diterapkan dengan menggunakan *node-node* yang terkait dengan metode *Hierarchical*. Adapun langkah - langkah dalam pembentukan *cluster* untuk membentuk suatu kelompok dalam *hierarchical cluster* yaitu:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - y_{jk})^2} \tag{4}$$

Keterangan:
d_{ij} : Jarak dari objke i serta objek j
P : Jumlah faktor dari klaster
x_{ik} : Data dari subjek i pada variable k
y_{jk} : Data dari subjek j pada variabel k

Dalam penerapannya, metode *agglomerative hierarchical cluster* sering diterapkan dikarenakan proses *cluster* memiliki sifat ilmiah. Berikut merupakan langkah algoritma *agglomerative*:
a. Pertama tentukan N klaster yang akan dibentuk, dimana N merupakan jumlah objek yang diteliti yang memiliki jarak $D = dij$
b. Mencari matriks jarak pasangan klaster yang paling dekat, selanjutnya menentukan matriks jarak.
c. Menggabungkan setiap klaster yang terbukti sangat dekat.
d. Lalu ulangi kembali tahapan 2 dan 3 sampai seluruh objek dapat membentuk klaster.

6) Analisis Hasil Perbandingan
Tahapan menganalisis hasil dari ke 3 metode algoritma di atas dengan menentukan mana hasil *clustering* yang paling baik. Evaluasi performa yang diterapkan ke 3 metode algoritma yaitu menggunakan *Silhouette Coefficient*. Nilai dari *silhouette* didefinisikan dengan *sil* (k) dihitung dengan menggunakan rata-rata dari *silhouette* pada *cluster*, rumusnya sebagai berikut

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

:

$$sil(c) = sil(k)\frac{1}{|k|}\sum_{i=1}^{k} sil(C_i) \quad (5)$$

Keterangan :
$sil(k)$: Nilai dari *silhouette*
$|k|$: Jumlah *cluster* k
$sil(C_i)$: Mean nilai *silhouette*

7) Kesimpulan
Penjabaran hasil kesimpulan dari proses analisis *clustering* dengan ke 3 metode yakni K-Means, DBSCAN, *Hierarchical* beserta hasil perbandingannya.

III. Hasil dan Pembahasan

A. Data Segmentasi

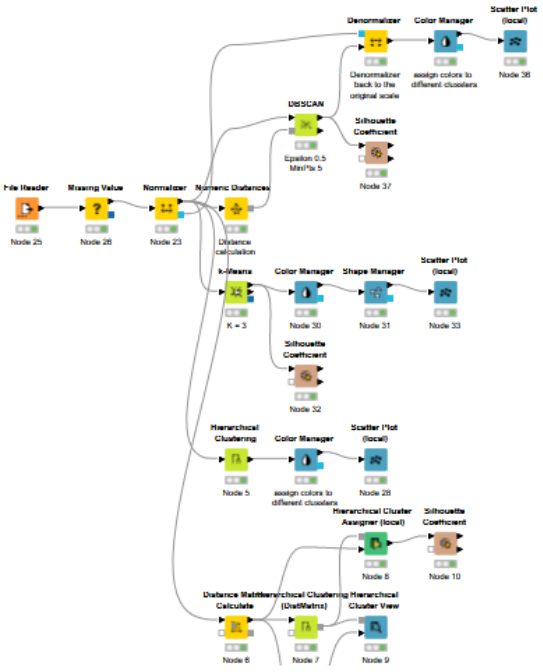
Data yang diterapkan pada penelitian ini yakni data mengenai segmentasi pasar penjualan otomotif yang diambil pada sumber kaggle.com dengan jumlah data sebanyak 2824 data. Data set yang digunakan ini terdapat 25 atribut, yakni sebagai berikut: *Order Number*, *Quantity Ordered*, *Price Each*, *Orderline Number*, *Sales*, *Order Date*, *Status*, *Qtr_id*, *Month_id*, *Year_id*, *Product Line*, *Msrp*, *Productcode*, *Customername*, *Phone*, *Addressline1*, *Addressline2*, *City*, *State*, *Postalcode*, *Country*, *Territory* ,*Contact Last Name*, *Contact Firstname*, *Dealsize*. Berikut ini adalah Tabel I, data dari segmentasi pasar yang diambil 5 data teratas :

Tabel I: Data Set 5 Data Teratas

ORDERNUMBER	QUANTITYORDERED	...	PRODUCTLINE	CUSTOMERNAME	...	CITY	DEALSIZE
10107	30	...	Motorcycles	Land of Toys Inc.	...	NYC	Small
10103	26	...	Classic Cars	Baane Mini Imports	...	Stavern	Medium
10113	21	...	Trucks and Buses	Mini Gifts Distributors Ltd.	...	San Rafael	Medium
10121	50	...	Motorcycles	Reims Collectables	...	Reims	Large
10119	43	...	Planes	Salzburg Collectables	...	Salzburg	Medium

B. Model Workflow Clustering

Gambar 2 di bawah ini merupakan alur dari proses *clustering* menggunakan metode algoritma K-Means , DBSCAN, dan *Hierarchical* menggunakan *tools* KNIME.



Gambar 2: Model Workflow Clustering

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

Node yang digunakan yang pertama adalah *node File Reader* yang digunakan untuk membaca *data set*. Lalu selanjutnya terdapat *node Missing Value* yang digunakan untuk menangani nilai yang kosong pada tabel dan dilanjutkan dengan normalisasi dengan menggunakan *node Normalizer*. Data set segmentasi pasar penjualan otomotif terdapat beberapa atribut yang data nya tidak lengkap, yang mana atribut tersebut yakni *Address Line 2*, *State*, dan *Postal Code*, dapat. dilihat pada. Tabel II berikut ini.

Tabel II: Data Set Missing Value

Row ID	...	ADDRESSLINE2	CITY	STATE	POSTALCODE
Row3	...	?	Pasadena	CA	90003
Row4	...	?	San Fransisco	CA	?
Row5	...	?	Burlingame	CA	94217
Row6	...	?	Lille	?	59000
Row7	...	?	Bergen	?	N 5804

Tabel II diatas merupakan data set segmentasi pasar yang belum dilakukannya preprocessing data dengan menggunakan *node Missing Value*, sehingga terlihat simbol tanda tanya yang menandakan bahwa pada tribut tersebut data nya tidak lengkap atau kosong. Preprocessing data dilakukan supaya pada proses memodelkan metode *clustering* tidak terdapat kendala pada hasil *clustering*. Dalam proses pemodelan *clustering* setiap metode algoritma memiliki *node* yang berbeda-beda dan digunakan sesuai dengan fungsi dan kegunaanya. Berikut ini adalah deskripsi dari masing-masing *node* yang diterapkan pada penelitian ini dapat dilihat pada Tabel III berikut ini.

Tabel III: Node Clustering KNIME

No.	Simbol *Node*	Nama *Node*	Keterangan
1		*File Reader*	Membaca file dengan berbagai jenis format
2		*Missing Value*	Menangani nilai kosong pada sel tabel
3		*Normalizer*	Menormalisasikan semua nilai numerik di setiap kolom
4		*Numeric Distances*	Mendeskripsikan jarak di setiap kolom yang numerik seperti jarak pada Euclidean atau Manhattan
5		*Denormalizer*	Mendenormalisasikan data inputan berdasarkan parameter normalisasi PMML
6		*Distance Matrix Calculate*	Menghitung nilai jarak pada seluruh pasangan baris pada tabel input
7		*K-Means*	Menentukan jumlah cluster untuk menghasilkan pusat cluster
8		*DBSCAN*	Menentukan nilai cluster dengan menggunakan MinPts dan Eps atau jarak tertentu
9		*Hierarchical Clustering*	Mengelompokkan data yang diinputkan
10		*Hierarchical Clustering (DistMatrix)*	Mengelompokkan data yang diinputkan dengan menggunakan matriks jarak

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

Tabel III – Lanjutan halaman sebelumnya

No.	Simbol *Node*	Nama *Node*	Keterangan
11		*Color Manager*	Menentukan warna pada atribut yang dipilih
12		*Shape Manager*	Menentukan simbol pada atribut yang dipilih
13		*Scatter Plot*	Menampilkan visualisasi scatter-plot dengan atribut yang dipilih
14		*Hierarchical Cluster View*	Memvisualisasikan data pengelompokan dendogram cluster
15		*Silhouette Coefficient*	Menghitung nilai koefisien silhouette pada cluster

C. Analisis Clustering dengan Metode K-MEANS

Proses pemodelan *clustering* metode algoritma K-Means dilakukan menggunakan *node* K-Means dengan melakukan pengclusteran sebanyak lima kali dimulai dari$k = 2$ sampai $k = 6$. Konfigurasi *node Normalizer* pada KNIME dilakukan dengan menyetting *Min-Max Normalizer*, dimana$Min = 0.0$ dan $Max = 1.0$. Mekanismenya adalah dengan mengurangkan setiap nilai fitur dengan nilai min fitur tersebut serta membaginya dengan rentang atau nilai max dikurangi nilai min fitur pada tersebut. Berikut ini rumus perhitungan dari *Min-Max Normalizer* :

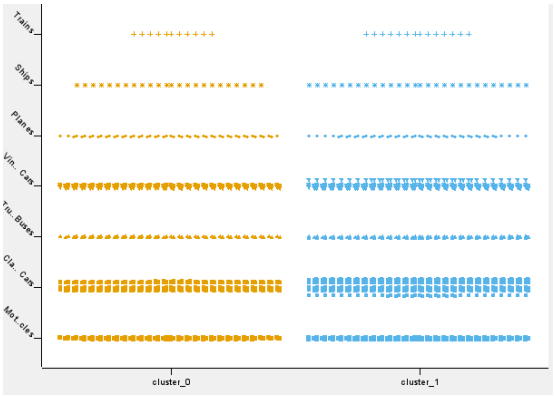
$$X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}} \tag{6}$$

Untuk menentukan nilai *cluster* terbaik dilakukan menggunakan *node Silhouette Coefficient* dengan menentukan *mean* dari *Silhouette Coefficient*. Hasil *clustering* metode. K-Means dapat dilihat pada Tabel IV dibawah ini.

Tabel IV: Nilai Cluster Rata-Rata *Silhouette Coefficient* K-Means

Row ID	Mean Silhouette Coefficient
2	0.716
3	0.464
4	0.496
5	0.511
6	0.625

Berdasarkan Tabel IV diatas nilai *cluster* yang terbaik adalah *cluster* $k = 2$, karena didapatkan nilai rata-rata *Silhouette Coefficient* nya tinggi yakni sebesar 0.716 dibandingkan *cluster* yang lain. Hasil visualisasi K-Means dengan *cluster* $k = 2$ dilakukan dengan menggunakan *node Color Manage*r, S*hape Manager*, dan *Scatterplot* dapat. dilihat pada Gambar 3.



Gambar 3: Visualisasi Scatter Plot k=2 K-Means

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

Proses visualisasi *Scatterplot* dilakukan dengan menggunakan *node Color Manager*, *Shape Manager*. Dapat dilihat pada garis *horizontal* terdapat keterangan *cluster_0* dan *cluster_1* yang merupakan nilai *cluster* yang diinputkan pada KNIME, sedangkan pada garis vertikal menunjukkan keterangan dari produk penjualan otomotif yang dikelompokkan, dimana produk penjualan terdiri dari *Motorcycles*, *Classic Cars*, *Trucks and Buses*, *Vintage Cars*, *Planes*, *Ships* dan *Trains*. Berdasarkan Scatterplot pada Gambar 3 di atas dapat diketahui bahwasannya setiap *cluster* dapat dilihat berdasarkan warna masing-masing dimana setiap *cluster* memiliki daerah persebaran yang berbeda-beda.

D. Analisis Clustering dengan Metode DBSCAN

Proses pemodelan *clustering* metode algoritma DBSCAN dilakukan menggunakan *node* DBSCAN dengan melakukan pengclusteran sebanyak lima kali dimulai dari $k = 2$ sampai$k = 6$. Sama halnya dengan metode K-MEANS konfigurasi Normalizer di *setting Min-Max Normalizer*. Konfigurasi *node Numeric Distance* pada KNIME dilakukan untuk menentukan jarak dengan *Euclidean Distance* dari data yang ada. Berikut ini rumus perhitungan dari *Euclidean Distance* :

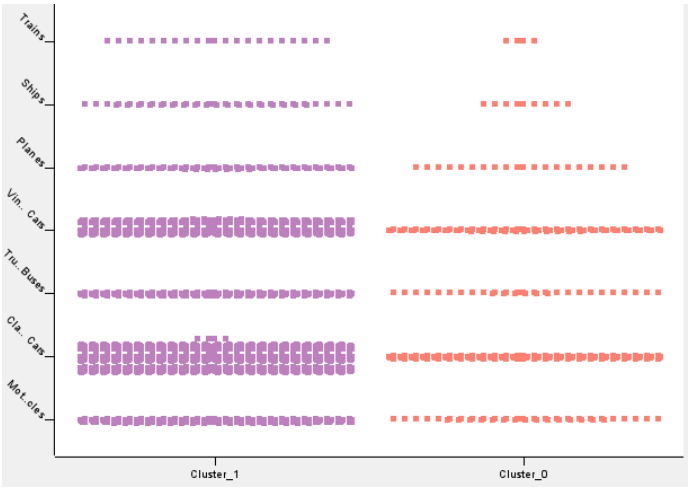
$$d = \sqrt{(x_1 - x_2)^2 + (c_1 - c_2)^2} \tag{7}$$

Konfigurasi pada *node* DBSCAN dilakukan penginputan nilai *Epsilon* dan MinPts. Untuk menentukan nilai *cluster* terbaik dilakukan menggunakan *node Silhouette Coefficient* dengan menentukan rata-rata dari *Silhouette Coefficient*. Hasil dari *clustering* metode DBSCAN dapat dilihat pada Tabel V dibawah ini.

Tabel V: Nilai Cluster Rata-Rata Silhouette Coefficient DBSCAN

Cluster	Minimum Points	Epsilon	Mean Silhouette Coefficient
2	5	0.529	0.298
3	5	0.5	-0.031
4	2	0.3499	-0.044
5	3	0.349	-0.023
6	3	0.3489	-0.006

Berdasarkan Tabel V diatas nilai *cluster* yang terbaik adalah *cluster* k = 2, karena didapatkan nilai rata-rata *Silhouette Coefficient* nya tinggi yakni sebesar 0.298 dengan MinPts 5 dan Eps 0.529 dibandingkan *cluster* yang lain. Hasil visualisasi DBSCAN dengan *clusterk* $= 2$ dilakukan dengan menggunakan *node Color Manager*, *Shape Manage*r, dan *Scatterplot* dapat dilihat pada Gambar 4 dibawah ini.



Gambar 4: Visualisasi Scatter Plot k=2 DBSCAN

Proses visualisasi Scatterplot dilakukan dengan menggunakan *node Color Manager*, *Shape Manager*. Berdasarkan *Scatterplot* pada Gambar 4 di atas dapat diketahui bahwasannya setiap *cluster* dapat dilihat berdasarkan warna masing-masing dimana setiap *cluster* memiliki daerah persebaran yang berbeda-beda.

E. Analisis Clustering dengan Metode Hierarchical

Proses pemodelan *clustering* metode algoritma *Hierarchical* dilakukan menggunakan *node Hierarchical* dengan melakukan pengclusteran sebanyak lima kali dimulai dari $k = 2$ sampai $k = 6$. Konfigurasi *node Distance Matriks Calculate* dilakukan dengan menentukan jarak *Euclidean Distance*. Selanjutnya dilakukan analisis dengan node berbeda berdasarkan algoritma

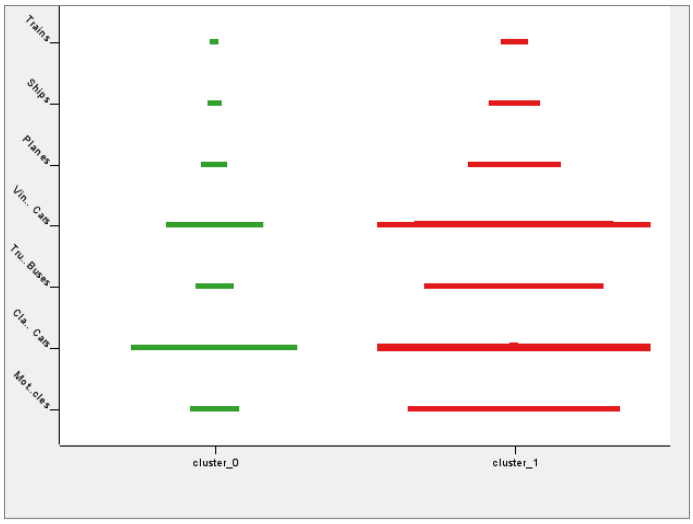
JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

terkait. Untuk menentukan nilai *cluster* terbaik dilakukan menggunakan *node Silhouette Coefficient* dengan menentukan rata-rata dari *Silhouette Coefficient*. Hasil dari *clustering metode Hierarchical* dapat dilihat. pada Tabel VI.

Tabel VI: Nilai Cluster Rata-Rata Silhouette Coefficient Hierarchical

Row ID	Mean Silhouette Coefficient
2	0.301
3	0.215
4	0.187
5	0.12
6	0.061

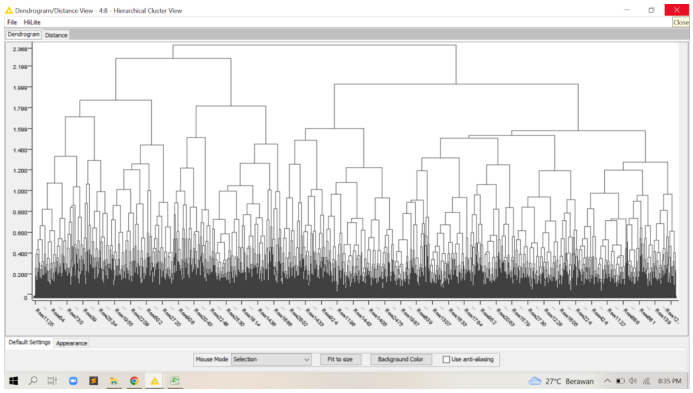
Berdasarkan Tabel VI diatas diketahui bahwa nilai *cluster* yang terbaik adalah k2 dengan nilai rata-rata sebesar 0.301 dimana nilai tersebut merupakan nilai rata-rata *Silhouette* tertinggi diantara nilai rata-rata *cluster* yang lain. Hasil visualisasi k2 dapat dilihat pada Gambar 5.



Gambar 5: Visualisasi Scatter Plot k=2 Hierarchical

Berdasarkan *Scatterplot* pada Gambar 5 di atas dapat diketahui bahwasannya setiap *cluster* dapat dilihat berdasarkan warna masing-masing dimana setiap *cluster* memiliki daerah persebaran yang berbeda-beda. Gambaran visual pada tahapan pada analisis *cluster* dalam menampilkan terbentuknya *cluster* serta nilai dari koefisien jarak dapat dilihat melalui Dendogram.

Gambar 6 di bawah ini merupakan hasil dendogram menggunakan *tools Hierarchical cluster view* pada Knime. Memastikan jarak setiap *cluster* digunakan dengan metode *single linkage* dimana dapat mengetahui jarak dua *cluster* yang ada lalu dipilih jarak terdekat. Angka sebelah kanan yang terdapat di dalam dendogram merupakan objek. penelitian, yang mana setiap objek direlasikan atas garis bersama objek yang lain, yang mana akan membuat satu *cluster*.



Gambar 6: Visualisasi Dendogram

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 191–201
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.796

F. Analisis Perbandingan 3 Metode Clustering

Hasil *clustering* dataset segmentasi pasar penjualan otomotif menggunakan metode algoritma K-Means, DBSCAN, dan *Hierarchical* dengan lima kali pengclusteran diperoleh nilai *cluster* terbaik pada masing-masing metode menggunakan *Silhouette Coefficient* tersaji dalam Tabel VII.

Tabel VII: Perbandingan Hasil Rata-Rata *Silhouette Coefficient*

Metode Clustering	Mean Silhouette Coefficient				
	k=2	k=3	k=4	k=5	k=6
K-Means	0.716	0.464	0.496	0.511	0.625
DBSCAN	0.298	-0.031	-0.044	-0.023	-0.006
Hierarchical	0.301	0.215	0.187	0.12	0.061

Dari proses *clustering* hasil dari pengelompokkan K-Means, DBSCAN, dan *Hierarchical* dicari nilai efektifitas serta kinerjanya dengan menggunakan *Silhouette Coefficient*, yang mana *Silhouette Coefficient* ini mengukur rata-rata jarak setiap titik pada *cluster*. Apabila nilai dari *Mean Silhouette Coefficient* mendekati angka 1 maka menandakan nilai hasil yang baik, namun jika angka mendekati -1 maka menandakan nilai hasilnya kurang baik. Hal tersebut dibuktikan oleh penelitian-penelitian sebelumnya yang membandingkan tentang beberapa metode *cluster* yang ada. Penelitian yang pernah dilakukan (Saputra dan Chusyairi, 2020) menjelaskan jika nilai *Silhouette Coefficient* mendekati 1 maka struktur *clustering* yang dihasilkan tepat, namun jika -1 sehingga struktur *clustering* yang diperoleh *overlapping* [21].

Pada Tabel VII diatas hasil *cluster* terbaik metode K-Means didapatkan pada *cluster* k =2 diperoleh nilai *mean Silhouette Coefficient* 0.178. Metode DBSCAN *cluster* yang cukup baik didapatkan pada *cluster* k = 2 dengan nilai *mean Silhouette Coefficient* 0.298, MinPts 5 dan Eps 0.529. Sedangkan metode *Hierarchical* didapatkan pada *cluster* k = 2 dengan nilai *mean Silhouette Coefficien*t 0.301. Berdasarkan pemahaman mengenai *Silhouette Coefficient* yang telah dipaparkan pada penelitian sebelumnya, dapat dilihat bahwasannya nilai *Silhouette Coefficient* dari setiap algoritma dimana metode algoritma K-Means memiliki kinerja yang lebih baik dari algoritma DBSCAN dan *Hierarchical*. Hal ini karena nilai *Mean Silhouette Coefficient* dari K-Means lebih besar sehingga mendekati 1, sedangkan rata-rata nilai *Silhouette Coefficient* DBSCAN lebih kecil dari K-Means, begitu juga dengan rata-rata nilai *Silhouette Coefficient Hierarchical* yang mendekati -1 dimana kinerjanya kurang baik. Hal ini menandakan bahwa metode algoritma K-Means memiliki kemampuan kinerja yang baik dalam melakukan *clustering* data.

IV. Simpulan

Hasil penelitian *clustering* dataset segmentasi pasar penjualan otomotif dengan K-Means diperoleh hasil *cluster* dengan performa kinerja yang baik yakni k =2 dengan nilai *Silhouette Coefficient* yang tinggi . Metode Algoritma K-Means memiliki performa kinerja yang baik dalam menganalisis data segmentasi pasar penjualan otomotif. Hal ini dikarenakan perolehan nilai *Silhouette Coefficient* yang cukup tinggi 0.716 sehingga dapat dikatakan mendekati nilai 1 dibandingkan dengan metode DBSCAN diperoleh 0.298 dengan MinPts 5, Eps 0.529 dan *Hierarchical* yang rendah. Waktu eksekusi pemodelan KNIME terbilang cukup singkat pada metode K-Means dibandingkan dengan metode DBSCAN dan *Hierarchical* yang lama.

Pustaka

[1] E. Fitriani, A. Saepudin, D. Ardiansyah, and R. Aryanti, "Implementasi Metode Naive Bayes Dalam Penyeleksian Karyawan untuk Penempatan Bagian Pemasaran," vol. 8, p. 7, 2022.
[2] N. Normah, S. Nurajizah, and A. Salbinda, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Fashion Hijab Banten," J. Tek. Komput., vol. 7, no. 2, pp. 158–163, Jul. 2021, doi: 10.31294/jtk.v7i2.10553.
[3] M. Rusdi, "Strategi Pemasaran untuk Meningkatkan Volume Penjualan pada Perusahaan Genting UD. Berkah Jaya," J. Studi Manaj. Dan Bisnis, vol. 6, no. 2, pp. 83–88, Dec. 2019, doi: 10.21107/jsmb.v6i2.6686.
[4] A. M. Afrilia, "Digital Marketing Sebagai Strategi Komunikasi Pemasaran 'Waroenk Ora Umum' Dalam Meningkatkan Jumlah Konsumen," J. Ris. Komun., vol. 1, no. 1, pp. 147–157, Feb. 2018, doi: 10.24329/jurkom.v1i1.21.
[5] H. Wijaya and H. Sirine, "Strategi Segmenting, Targeting, Positioning Serta Strategi Harga Pada Perusahaan Kecap Blekok Di Cilacap," AJIE, vol. 1, no. 3, pp. 175–190, Sep. 2016, doi: 10.20885/ajie.vol1.iss3.art2.
[6] N. Normah, B. Rifai, and P. Sari, "Algoritma Apriori Sebagai Solusi Kontrol Persediaan Suku Cadang Mobil PT. Buanasakti Aneka Motor Jakarta," Paradigma, vol. 22, no. 2, Art. no. 2, Sep. 2020, doi: 10.31294/p.v22i2.6530.
[7] J. Nasir, "Penerapan Data Mining Clustering Dalam Mengelompokan Buku Dengan Metode K-Means," Simetris J. Tek. Mesin Elektro Dan Ilmu Komput., vol. 11, no. 2, Art. no. 2, 2020, doi: 10.24176/simet.v11i2.5482.
[8] S. Nurdiani, S. Linawati, R. A. Safitri, and E. P. Saputra, "Pengelompokan Perilaku Mahasiswa Pada Perkuliahan E-Learning dengan K-Means Clustering," vol. 19, no. 2, p. 8, 2019.
[9] H. Priyatman, F. Sajid, and D. Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa," J. Edukasi Dan Penelit. Inform. JEPIN, vol. 5, no. 1, p. 62, Apr. 2019, doi: 10.26418/jp.v5i1.29611.
[10] H. Haviluddin, S. J. Patandianan, G. M. Putra, N. Puspitasari, and H. S. Pakpahan, "Implementasi Metode K-Means Untuk Pengelompokkan Rekomendasi Tugas Akhir," Inform. Mulawarman J. Ilm. Ilmu Komput., vol. 16, no. 1, p. 13, Mar. 2021, doi: 10.30872/jim.v16i1.5182.
[11] B. S. Ashari, S. C. Otniel, and R. Rianto, "Perbandingan Kinerja K-Means Dengan Dbscan Untuk Metode Clustering Data Penjualan Online Retail," J. Siliwangi Seri Sains Dan Teknol., vol. 5, no. 2, Art. no. 2, Dec. 2019.
[12] M. H. Adiya and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru," vol. 05, no. 01, p. 8, 2019.
[13] R. K. Dinata, N. Hasdyna, and N. Azizah, "Analisis K-Means Clustering pada Data Sepeda Motor," vol. 5, no. 1, p. 8, 2020.

[14] B. N. Sari and A. Primajaya, "Penerapan Clustering Dbscan Untuk Pertanian Padi Di Kabupaten Karawang," JIKO J. Inform. Dan Komput., vol. 4, no. 1, Art. no. 1, Sep. 2019, doi: 10.26798/jiko.v4i1.178.
[15] A. M. Sikana and A. W. Wijayanto, "Analisis Perbandingan Pengelompokan Indeks Pembangunan Manusia Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarchical Clustering," J. Ilmu Komput., vol. 14, no. 2, Art. no. 2, Sep. 2021, doi: 10.24843/JIK.2021.v14.i02.p01.
[16] D. Mulyaningrum and M. Nusrang, "Analisis Cluster Pendekatan Metode Hierarchical Clustering Terhadap Pertumbuhan Ekonomi Di Provinsi Sulawesi Selatan," p. 9.
[17] S. Suhirman and H. Wintolo, "System for Determining Public Health Level Using the Agglomerative Hierarchical Clustering Method," Compiler, vol. 8, no. 1, Art. no. 1, Mar. 2019, doi: 10.28989/compiler.v8i1.425.
[18] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coeficient Untuk Evaluasi Clutering Obat," vol. 6, no. 2, p. 7, 2021.
[19] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," Techno.Com, vol. 20, no. 2, Art. no. 2, May 2021, doi: 10.33633/tc.v20i2.4556.
[20] S. S. Alzahrani, "Data Mining Regarding Cyberbullying in the Arabic Language on Instagram Using KNIME and Orange Tools," Eng. Technol. Appl. Sci. Res., vol. 12, no. 5, pp. 9364–9371, Oct. 2022, doi: 10.48084/etasr.5184.
[21] Pelsri Ramadar Noor Saputra and A. Chusyairi, "Perbandingan Metode Clustering dalam Pengelompokan Data Puskesmas pada Cakupan Imunisasi Dasar Lengkap," J. RESTI Rekayasa Sist. Dan Teknol. Inf., vol. 4, no. 6, Dec. 2020, doi: 10.29207/resti.v4i6.2556.

