***JSAI : Journal Scientific and Applied Informatics***
Vol. 09, No. 2, *June* 2026, *pages*. 294~305
E-ISSN: 2614-3054; P-ISSN: 2614-3062, *accredited by Kemenristekdikti*, Sinta 4
DOI: 10.36085

# Perbandingan *K-Means, Hierarchical Clustering* Dan *K-Medoids* Untuk Segmentasi Pasar Berdasarkan Evaluasi *Silhouette Score*

**[1]Denny Ganjar Purnama, [2]Safrizal, [3]Cahyono Budy Santoso**
[1,2,3]Universitas Pembangunan Jaya, Indonesia
[1]*denny.ganjar@upj.ac.id*; [2]*safrizal.abdurrahman@upj.ac.id*; [3]*cahyono.budy@upj.ac.id*

**Article Info**

***Article history:***

Received, 2026-06-22
Revised, 2026-06-27
Accepted, 2026-06-30

***Kata Kunci:***

*K-Means*
*Hierarchical Clustering*
*K-Medoids*
*Silhouette Score*

**ABSTRAK**

Penelitian ini bertujuan membandingkan kinerja algoritma *K-Means*, *Agglomerative Hierarchical Clustering*, dan *K-Medoids* untuk segmentasi pasar berdasarkan data penjualan PT XYZ. Dataset menggunakan atribut *Quantity* dan *Expected Revenue* yang diproses melalui pembersihan data, konversi nilai mata uang ke numerik, penghapusan data tidak valid, transformasi logaritmik, dan standardisasi sehingga diperoleh 923 data valid. Proses *clustering* dilakukan dengan tiga klaster yang diinterpretasikan sebagai *Low Value, Mid Value,* dan *High Value*. Evaluasi menggunakan *Silhouette Score* menunjukkan bahwa *K-Means* memperoleh nilai 0,4805, *Agglomerative Hierarchical Clustering* 0,4808, dan *K-Medoids* 0,4840. Meskipun selisih nilai antaralgoritma relatif kecil, *K-Medoids* menghasilkan skor tertinggi sehingga dipilih sebagai model akhir. Hasil segmentasi *K-Medoids* terdiri atas 188 objek (20,37%) pada segmen *Low Value*, 469 objek (50,81%) pada segmen *Mid Value*, dan 266 objek (28,82%) pada segmen *High Value*. Temuan ini menunjukkan bahwa *K-Medoids* mampu menghasilkan kualitas klaster terbaik sekaligus memberikan interpretasi yang lebih mudah melalui pusat klaster berbasis objek aktual (*medoid*). Hasil penelitian dapat dimanfaatkan sebagai dasar penyusunan strategi pemasaran yang berbeda untuk setiap segmen pelanggan.

***Keywords:***

*K-Means*
*Hierarchical Clustering*
*K-Medoids*
*Silhouette Score*

***ABSTRACT***

*This study compares the performance of K-Means, Agglomerative Hierarchical Clustering, and K-Medoids algorithms for market segmentation using PT XYZ sales data. The dataset consists of the Quantity and Expected Revenue attributes and was processed through data cleaning, currency-to-numeric conversion, invalid data removal, logarithmic transformation, and standardization, resulting in 923 valid records. Clustering was performed using three clusters, representing Low Value, Mid Value, and High Value customer segments. Performance was evaluated using the Silhouette Score, where K-Means achieved 0.4805, Agglomerative Hierarchical Clustering 0.4808, and K-Medoids 0.4840. Although the performance differences among the algorithms were relatively small, K-Medoids achieved the highest score and was therefore selected as the final model. The resulting segmentation consisted of 188 Low Value customers (20.37%), 469 Mid Value customers (50.81%), and 266 High Value customers (28.82%). These findings indicate that K-Medoids provides the best clustering quality while offering greater interpretability through medoid-based cluster centers representing actual data objects. The proposed segmentation can support companies in developing differentiated marketing strategies for low-, medium-, and high-value customer segments.*



***Penulis Korespondensi:***

Denny Ganjar Purnama,
Program Studi Sistem Informasi,
Universitas Pembangunan Jaya,
Email: denny.ganjar@upj.ac.id

## 1. PENDAHULUAN

Perkembangan ekonomi digital dan meningkatnya intensitas persaingan bisnis mendorong perusahaan untuk mengelola informasi penjualan secara lebih cermat, terukur, dan berbasis data. Dalam konteks pemasaran modern, data penjualan tidak lagi dipandang semata-mata sebagai catatan transaksi, melainkan sebagai sumber pengetahuan strategis yang dapat digunakan untuk memahami perilaku pelanggan, kontribusi pendapatan, pola permintaan, serta peluang pertumbuhan pasar. Perusahaan yang memiliki ragam produk dan melayani lebih dari satu jenis pelanggan, seperti sektor privat dan sektor pemerintah/negeri, membutuhkan mekanisme segmentasi yang mampu membedakan karakteristik transaksi secara akurat. Segmentasi pasar menjadi penting karena pelanggan tidak memiliki kebutuhan, daya beli, frekuensi pembelian, dan nilai transaksi yang seragam. Kajian mengenai segmentasi pasar menegaskan bahwa pengelompokan pasar dapat membantu perusahaan menentukan target, menyusun posisi produk, dan meningkatkan *volume* penjualan melalui strategi yang lebih tepat sasaran [1].

Dalam praktik bisnis, sebagian perusahaan masih melakukan evaluasi penjualan dengan cara konvensional, misalnya hanya melihat total penjualan, jumlah unit terjual, atau daftar pelanggan dengan kontribusi pendapatan terbesar. Pendekatan seperti ini memang mudah dilakukan, tetapi kurang memadai untuk menggambarkan struktur pasar yang kompleks. Data penjualan PT XYZ, misalnya, memuat beberapa atribut seperti *Opportunity Name*, *Model Name*, *Product Line*, *Quantity*, *Probability*, *Expected Revenue*, dan *Created Date*. Atribut tersebut dapat menyimpan pola tersembunyi mengenai nilai transaksi, volume pembelian, dan kecenderungan sektor pelanggan. Tanpa pendekatan analitik yang memadai, perusahaan berisiko memperlakukan seluruh pelanggan dengan strategi yang sama, padahal masing-masing kelompok pelanggan dapat membutuhkan pendekatan pemasaran, layanan, dan prioritas penjualan yang berbeda.

Data mining menyediakan pendekatan sistematis untuk mengubah data mentah menjadi informasi yang bernilai bagi pengambilan keputusan. Data mining berperan dalam mengungkap pola, tren, hubungan tersembunyi, serta struktur data yang tidak selalu dapat dilihat melalui inspeksi manual [2]. Salah satu teknik data mining yang relevan untuk masalah segmentasi pasar adalah *clustering*, yaitu proses pengelompokan objek ke dalam beberapa kelompok berdasarkan kemiripan karakteristiknya. *Clustering* dapat membantu organisasi menemukan kelompok pelanggan yang serupa tanpa perlu label kelas di awal, sehingga cocok digunakan pada kasus segmentasi pasar berbasis data penjualan. Penelitian tentang penerapan *clustering* menunjukkan bahwa metode ini dapat digunakan untuk mengelompokkan calon siswa, pelanggan, maupun data transaksi ke dalam kelompok yang lebih mudah dianalisis [3].

*K-Means* merupakan salah satu algoritma *clustering* partisional yang banyak digunakan karena konsepnya sederhana, komputasinya relatif efisien, dan hasilnya mudah diinterpretasikan. Algoritma ini bekerja dengan menentukan sejumlah klaster, menginisialisasi *centroid*, menghitung jarak objek ke *centroid*, menempatkan objek pada klaster terdekat, lalu memperbarui *centroid* sampai kondisi konvergen [4]. Keunggulan tersebut membuat *K-Means* banyak dipakai dalam analisis bisnis dan segmentasi penjualan. Penelitian pada data penjualan supermarket, misalnya, menunjukkan bahwa *K-Means* dapat mengelompokkan cabang usaha berdasarkan karakteristik penjualan dan memberikan dukungan analitis bagi manajemen cabang [5]. Penelitian lain pada sentra UMKM juga menunjukkan bahwa *K-Means* dapat membantu membentuk segmen pelanggan yang lebih informatif bagi strategi pemasaran [6].

Di sisi lain, *Hierarchical Clustering* menawarkan pendekatan yang berbeda karena membentuk struktur pengelompokan secara bertingkat. Pada pendekatan *agglomerative*, setiap objek awalnya diperlakukan sebagai klaster tunggal, kemudian klaster-klaster yang paling mirip digabungkan secara bertahap hingga terbentuk struktur dendrogram. Keunggulan metode ini terletak pada kemampuannya menampilkan hubungan antarobjek dan antarklaster secara visual melalui pohon hierarki [7]. Dalam kasus segmentasi pelanggan, *Agglomerative Hierarchical Clustering* telah digunakan untuk mengelompokkan pelanggan barbershop berdasarkan perilaku transaksi dan dievaluasi menggunakan *Silhouette Coefficient* [8]. Dengan demikian, *Hierarchical Clustering* layak dibandingkan dengan *K-Means*, khususnya untuk menilai apakah data penjualan memiliki struktur segmentasi alami yang konsisten.

*K-Medoids* merupakan algoritma *clustering* partisional yang memiliki prinsip mirip dengan *K-Means*, tetapi pusat klasternya bukan *centroid* hasil rata-rata, melainkan medoid atau objek aktual yang paling representatif dalam klaster. Perbedaan ini penting pada data penjualan karena nilai transaksi sering kali memiliki sebaran tidak seimbang dan mengandung nilai ekstrem. Pada kondisi tersebut, *medoid* dapat memberikan representasi yang lebih mudah dijelaskan secara bisnis karena pusat segmen berupa pelanggan atau transaksi nyata, bukan titik rata-rata yang mungkin tidak terdapat dalam dataset [9], [10]. Oleh sebab itu, *K-Medoids* layak ditambahkan sebagai algoritma pembanding untuk menilai apakah segmentasi yang dihasilkan lebih stabil dibandingkan *K-Means* dan *Hierarchical Clustering*.

Perbandingan algoritma menjadi penting karena tidak ada satu metode *clustering* yang selalu unggul untuk semua bentuk data. Kualitas hasil *clustering* dipengaruhi oleh jumlah atribut, bentuk distribusi data, skala fitur, keberadaan *outlier*, kepadatan antarkelompok, serta metode penentuan jumlah klaster. Penelitian komparatif pada data segmentasi pasar menunjukkan bahwa *K-Means*, DBSCAN, dan *Hierarchical* dapat menghasilkan performa berbeda ketika dievaluasi menggunakan *Mean Silhouette Coefficient* [11]. Penelitian lain yang membandingkan *Hierarchical*, *K-Means*, dan DBSCAN pada data penjualan melalui *Facebook* juga menegaskan bahwa karakteristik dataset memengaruhi kualitas kelompok yang terbentuk [12]. Oleh karena itu, perbandingan antara *K-Means* dan *Hierarchical Clustering* pada data penjualan PT XYZ menjadi relevan untuk menentukan algoritma yang paling sesuai bagi kebutuhan segmentasi pasar perusahaan.

Berdasarkan uraian tersebut, penelitian ini difokuskan pada perbandingan *K-Means, Hierarchical Clustering*, dan *K-Medoids* untuk segmentasi pasar berbasis data penjualan PT XYZ dengan evaluasi *Silhouette Score*. Penelitian ini memiliki empat kontribusi utama. Pertama, penelitian menyusun prosedur segmentasi pasar berbasis data mining dengan kerangka CRISP-DM sehingga tahapan analisis dapat ditelusuri secara sistematis. Kedua, penelitian membandingkan tiga algoritma *clustering*, yakni *K-Means*, *Agglomerative Hierarchical Clustering*, dan *K-Medoids*, pada dataset yang sama, atribut yang sama, dan jumlah klaster yang sama. Ketiga, penelitian menambahkan perspektif *robustness* melalui *K-Medoids* karena pusat klaster dipilih dari objek aktual atau medoid, bukan rata-rata *centroid* sebagaimana pada *K-Means* [10], [13], [14]. Keempat, penelitian menghubungkan hasil *clustering* dengan interpretasi bisnis dalam bentuk segmen *Low Value*, *Mid Value*, dan *High Value* serta distribusi sektor privat dan pemerintah/negeri. Dengan demikian, hasil penelitian tidak hanya memberikan nilai teknis berupa skor validitas klaster, tetapi juga memberikan wawasan manajerial bagi perusahaan untuk mendukung pengambilan keputusan pemasaran.

Rumusan masalah dalam penelitian ini adalah bagaimana hasil segmentasi pasar berdasarkan data penjualan PT XYZ menggunakan *K-Means*, *Hierarchical Clustering*, dan *K-Medoids*, serta algoritma mana yang memberikan kualitas klaster lebih baik berdasarkan *Silhouette Score*. Adapun tujuan penelitian adalah menganalisis karakteristik data penjualan berdasarkan atribut *Quantity* dan *Expected Revenue*, menentukan jumlah klaster optimal, membentuk segmentasi pasar menggunakan *K-Means, Hierarchical Clustering*, dan *K-Medoids*, membandingkan hasil evaluasi ketiga algoritma, serta merumuskan interpretasi bisnis dari segmen yang terbentuk. Penelitian ini diharapkan dapat memberikan manfaat akademik sebagai referensi penerapan *clustering* pada data penjualan dan manfaat praktis bagi perusahaan dalam menyusun strategi pemasaran berbasis segmentasi data.

## 2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan jenis penelitian komparatif. Pendekatan kuantitatif digunakan karena objek yang dianalisis berupa data numerik penjualan, sedangkan pendekatan komparatif digunakan untuk membandingkan kualitas hasil *clustering* antara *K-Means*, *Agglomerative Hierarchical Clustering*, dan *K-Medoids*. Fokus penelitian bukan untuk memprediksi kelas pelanggan, melainkan untuk menemukan struktur kelompok yang tersembunyi dalam data penjualan. Oleh karena itu, metode yang digunakan termasuk dalam pembelajaran tidak terawasi atau unsupervised learning. Hasil akhir penelitian diarahkan pada pembentukan segmen pasar serta evaluasi kualitas klaster menggunakan Silhouette Score.

Kerangka kerja penelitian mengacu pada *Cross Industry Standard Process for Data Mining* atau CRISP-DM. CRISP-DM dipilih karena menyediakan tahapan analisis data mining yang sistematis, mulai dari pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, hingga penyajian hasil [15]. Dalam penelitian ini, CRISP-DM digunakan untuk memastikan bahwa pemodelan *clustering* tidak dilakukan secara langsung tanpa memahami konteks bisnis dan kondisi data. Tahap business *understanding* digunakan untuk merumuskan kebutuhan segmentasi pasar. Tahap data understanding digunakan untuk memahami struktur dataset, atribut, dan periode data. Tahap data preparation digunakan untuk membersihkan dan mentransformasi data. Tahap modeling digunakan untuk menerapkan *K-Means, Hierarchical Clustering*, dan *K-Medoids*. Tahap *evaluation* digunakan untuk membandingkan *Silhouette Score* ketiga algoritma. Tahap *deployment* diterjemahkan sebagai penyusunan interpretasi bisnis dari klaster yang terbentuk.

Tabel 1. Implementasi tahapan CRISP-DM

| Tahapan CRISP-DM | Implementasi dalam Penelitian |
|---|---|
| *Business Understanding* | Merumuskan kebutuhan segmentasi pasar PT XYZ berdasarkan data penjualan sektor privat dan pemerintah/negeri. |
| *Data Understanding* | Menganalisis struktur data, atribut transaksi, periode data, dan kelayakan fitur untuk *clustering*. |
| *Data Preparation* | Membersihkan nilai tidak valid, mengonversi *Expected Revenue* menjadi numerik, melakukan log *transformation*, dan standardisasi. |
| *Modeling* | Menerapkan *K-Means, Agglomerative Hierarchical Clustering*, dan *K-Medoids* dengan jumlah klaster tiga. |
| *Evaluation* | Membandingkan kualitas klaster ketiga algoritma menggunakan *Silhouette Score*. |
| *Deployment/Interpretasi* | Menerjemahkan klaster menjadi segmen *Low Value, Mid Value*, dan *High Value* serta menganalisis sektor pelanggan. |

Data penelitian diperoleh dari data penjualan PT XYZ yang diakses dari bagian marketing perusahaan. Dataset mentah yang digunakan dalam revisi ini adalah *file Sellout* 2022-jan 2026.xlsx. Berdasarkan kolom *Created Date*, data valid yang dianalisis mencakup rentang Juni 2021 sampai Februari 2026. Dataset awal memuat 987 baris data transaksi dan ringkasan, kemudian setelah proses pembersihan diperoleh 923 objek valid untuk proses *clustering*. Atribut yang tersedia mencakup *Opportunity Name, Model Name, Product Line, Quantity, Sales Price, Probability* (%), *Stage, Expected Revenue*, dan *Created Date*. Namun, atribut utama yang digunakan dalam pemodelan adalah *Quantity* dan *Expected Revenue* karena keduanya merepresentasikan volume pembelian dan kontribusi nilai transaksi. Atribut lain tetap dipertahankan sebagai informasi pendukung untuk interpretasi bisnis.

Tahap data preparation dilakukan untuk memastikan data siap diproses oleh algoritma berbasis jarak. Data penjualan memiliki format mata uang pada atribut Sales *Price* dan *Expected Revenue*, sehingga nilai tersebut perlu dikonversi menjadi numerik. Proses pembersihan juga dilakukan dengan menghapus baris kosong, baris ringkasan, nilai yang tidak dapat dikonversi, serta nilai *Quantity* atau *Expected Revenue* yang tidak valid. Dari 987 baris data awal, sebanyak 64 baris tidak digunakan karena tidak memenuhi kriteria validitas data, sehingga data akhir yang dianalisis berjumlah 923 objek. Tahap ini penting karena kualitas *preprocessing* sangat memengaruhi jarak antardata dan hasil klaster yang dihasilkan oleh *K-Means, Hierarchical Clustering*, maupun *K-Medoids* [16], [2].

Setelah proses pembersihan, dilakukan seleksi atribut. Seleksi atribut penting karena algoritma *clustering* berbasis jarak akan menghitung kedekatan antarobjek berdasarkan fitur yang digunakan. Jika fitur yang tidak relevan dimasukkan, hasil klaster dapat menjadi bias dan sulit diinterpretasikan. *Quantity* dipilih karena menunjukkan jumlah unit produk yang dibeli pelanggan, sedangkan *Expected Revenue* dipilih karena menunjukkan nilai moneter transaksi. Kedua atribut tersebut mencerminkan dua dimensi utama segmentasi penjualan, yaitu dimensi *volume* dan dimensi kontribusi pendapatan. Kombinasi keduanya memungkinkan perusahaan membedakan pelanggan yang membeli sedikit dengan nilai transaksi rendah, pelanggan yang membeli sedikit tetapi nilai transaksinya tinggi, dan pelanggan yang membeli banyak dengan nilai moneter tinggi.

Tabel 2. Seleksi atribut penelitian

| Atribut | Tipe Data | Peran dalam Penelitian |
|---|---|---|
| *Quantity* | Numerik | Merepresentasikan jumlah unit produk yang dibeli pelanggan dan menjadi indikator volume transaksi. |
| *Expected Revenue* | Numerik | Merepresentasikan nilai moneter transaksi dan menjadi indikator kontribusi pendapatan. |
| *Opportunity Name* | Kategorikal/teks | Tidak dihitung sebagai fitur jarak, tetapi digunakan untuk interpretasi sektor privat atau pemerintah/negeri. |
| *Model Name, Product Line, Probability, Created Date* | Kategorikal/administratif/waktu | Dipertahankan sebagai informasi pendukung, tetapi tidak dimasukkan ke perhitungan jarak utama. |

Pemilihan hanya dua atribut, yaitu *Quantity* dan *Expected Revenue*, dilakukan berdasarkan pertimbangan teoritis maupun praktis. Kedua atribut tersebut secara langsung merepresentasikan dimensi utama segmentasi pelanggan, yaitu volume transaksi dan kontribusi nilai ekonomi. Atribut lain seperti *Product Line*, *Opportunity Name*, *Probability*, dan *Created Date* tidak digunakan sebagai fitur utama karena sebagian besar bersifat kategorikal atau administratif sehingga kurang sesuai untuk algoritma *clustering* berbasis jarak *Euclidean* tanpa

transformasi tambahan. Penggunaan atribut yang terlalu banyak juga berpotensi menurunkan interpretabilitas hasil segmentasi dan menyebabkan noise pada proses pengelompokan.

Transformasi data dilakukan dalam dua tahap, yaitu log transformation dan standardisasi. *Log transformation* menggunakan fungsi log(1+x) atau np.log1p untuk mereduksi skewness dan menekan pengaruh nilai ekstrem pada atribut *Quantity* dan *Expected Revenue*. Transformasi ini diperlukan karena data penjualan biasanya memiliki sebaran tidak simetris, yaitu sebagian besar transaksi bernilai kecil sampai menengah dan sebagian kecil transaksi bernilai sangat besar. Setelah transformasi logaritmik, data distandardisasi menggunakan StandardScaler sehingga setiap atribut memiliki rata-rata 0 dan standar deviasi 1. Standardisasi diperlukan karena algoritma berbasis jarak sangat sensitif terhadap perbedaan skala antaratribut [17].

Pada tahap *modeling*, algoritma pertama yang digunakan adalah *K-Means*. Secara konseptual, *K-Means* membentuk k klaster dengan meminimalkan variasi dalam klaster. Prosesnya dimulai dengan menentukan jumlah klaster, memilih *centroid* awal, menghitung jarak setiap objek ke *centroid*, menempatkan objek pada *centroid* terdekat, lalu menghitung ulang *centroid* sampai tidak terjadi perubahan yang signifikan [4]. Dalam penelitian ini, jarak yang digunakan adalah *Euclidean Distance* karena kedua fitur telah distandardisasi dan berada pada skala yang sebanding. Jumlah klaster ditentukan menggunakan *Elbow Method*. Rumus umum SSE dapat ditulis sebagai SSE = sum dari seluruh jarak kuadrat objek terhadap *centroid* klasternya. Penurunan SSE yang tajam menunjukkan bahwa penambahan klaster meningkatkan homogenitas kelompok, sedangkan penurunan yang mulai melandai menunjukkan bahwa penambahan klaster berikutnya tidak lagi memberikan manfaat besar [18].

Algoritma kedua yang digunakan adalah *Agglomerative Hierarchical Clustering*. Metode ini dimulai dengan menganggap setiap objek sebagai klaster tunggal, kemudian menggabungkan dua klaster yang paling dekat secara bertahap sampai terbentuk struktur hierarki [7]. Penelitian ini menggunakan *Ward Linkage* karena metode *Ward* berupaya meminimalkan peningkatan total varians dalam klaster ketika dua klaster digabungkan. *Ward Linkage* dinilai cocok untuk data yang telah distandardisasi dan bertujuan menghasilkan kelompok yang relatif kompak [19], [20]. Hasil proses *Hierarchical* divisualisasikan menggunakan dendrogram. Jumlah klaster ditentukan dengan memotong dendrogram pada tingkat dissimilarity yang dianggap menghasilkan struktur kelompok paling representatif. Agar perbandingan adil, jumlah klaster pada *Hierarchical* diselaraskan dengan hasil *K-Means*, yaitu tiga klaster.

Algoritma ketiga yang digunakan adalah *K-Medoids. K-Medoids* termasuk metode partitional *clustering* yang membagi data ke dalam k kelompok, tetapi pusat kelompok dipilih dari objek nyata dalam dataset. Tujuan utama *K-Medoids* adalah meminimalkan total jarak antara setiap objek dan medoid terdekatnya. Secara umum, fungsi objektif *K-Medoids* dapat dituliskan sebagai minimisasi jumlah *dissimilarity* d(xi, mj), dengan xi sebagai objek data dan mj sebagai medoid dari klaster j. Karena medoid merupakan objek aktual, hasil *K-Medoids* lebih mudah ditelusuri untuk interpretasi bisnis, misalnya dengan melihat transaksi atau pelanggan yang menjadi wakil paling representatif dari segmen *Low Value, Mid Value*, dan *High Value* [9], [14].

Dalam penelitian ini, *K-Medoids* dijalankan pada dataset hasil transformasi yang sama dengan *K-Means* dan *Hierarchical Clustering*, yaitu fitur *Quantity* dan *Expected Revenue* yang telah melalui log transformation dan StandardScaler. Jumlah klaster yang digunakan juga sama, yaitu k = 3, agar hasil evaluasi dapat dibandingkan secara adil. Implementasi *K-Medoids* dilakukan dengan prosedur assignment-update berbasis *medoid*, yaitu memilih medoid awal, mengalokasikan objek ke *medoid* terdekat, kemudian memperbarui medoid dalam setiap klaster berdasarkan objek aktual yang meminimalkan total jarak intra-klaster. Prosedur ini menggunakan *random_state* = 42, *n_init* = 100, dan *max_iter* = 100 agar hasil lebih stabil dan replikatif.

Evaluasi dilakukan menggunakan *Silhouette Score*. Indeks ini menggabungkan dua konsep, yaitu *cohesion* dan *separation*. *Cohesion* menunjukkan seberapa dekat suatu objek dengan objek lain dalam klaster yang sama, sedangkan *separation* menunjukkan seberapa jauh objek tersebut dari klaster terdekat lainnya. Untuk setiap objek i, nilai *silhouette* dapat ditulis sebagai s(i) = (b(i) - a(i)) / max(a(i), b(i)), dengan a(i) sebagai rata-rata jarak objek i terhadap objek lain dalam klaster yang sama dan b(i) sebagai rata-rata jarak terendah objek i terhadap klaster lain. Nilai s(i) berada pada rentang -1 sampai 1. Nilai mendekati 1 menunjukkan objek berada pada klaster yang tepat, nilai mendekati 0 menunjukkan objek berada di batas antarklaster, dan nilai negatif menunjukkan objek kemungkinan lebih sesuai berada pada klaster lain. Penggunaan *Silhouette Coefficient* juga telah diterapkan dalam penelitian segmentasi pelanggan dan komparasi algoritma *clustering* sebelumnya [8], [11].

Hasil klaster kemudian diberi label bisnis. Label teknis yang dihasilkan algoritma, misalnya cluster 0, 1, dan 2, tidak langsung memiliki makna manajerial. Oleh karena itu, masing-masing klaster diurutkan berdasarkan rata-rata *Expected Revenue* dan *Quantity*. Klaster dengan rata-rata nilai terendah diberi label *Low Value*, klaster menengah diberi label *Mid Value*, dan klaster tertinggi diberi label *High Value*. Proses pelabelan ini penting agar hasil analitik dapat digunakan oleh manajemen pemasaran. Setelah label terbentuk, dilakukan analisis

distribusi jumlah pelanggan per segmen serta distribusi sektor privat dan pemerintah/negeri. Klasifikasi sektor dilakukan berdasarkan pola kata pada *Opportunity Name*, misalnya kata kunci yang mengarah pada kementerian, dinas, pemerintah, atau institusi negeri dikategorikan sebagai pemerintah/negeri, sedangkan sisanya dikategorikan sebagai privat. Alur pengolahan data dalam penelitian ini dimulai dari import dataset penjualan, pemeriksaan struktur atribut, pembersihan nilai tidak valid, seleksi fitur *Quantity* dan *Expected Revenue*, transformasi logaritmik, standardisasi, penentuan jumlah klaster, pemodelan *K-Means, Hierarchical Clustering*, dan *K-Medoids*, evaluasi *Silhouette Score*, serta interpretasi bisnis. Alur tersebut dibuat agar setiap tahap dapat ditelusuri dan direplikasi oleh peneliti selanjutnya.

Secara ringkas, prosedur *K-Means* terdiri atas: menentukan k, menginisialisasi *centroid*, menghitung jarak *Euclidean*, mengalokasikan objek ke *centroid* terdekat, memperbarui *centroid*, dan mengulang proses sampai konvergen. Prosedur *Hierarchical* terdiri atas: memperlakukan setiap objek sebagai klaster tunggal, menghitung jarak antarklaster, menggabungkan klaster dengan peningkatan varians terkecil berdasarkan Ward *Linkage*, membangun dendrogram, dan menentukan titik potong untuk memperoleh tiga klaster. Secara ringkas, prosedur *K-Medoids* terdiri atas: menentukan jumlah klaster k = 3, memilih medoid awal, menghitung jarak setiap objek terhadap *medoid*, menempatkan objek ke medoid terdekat, mengevaluasi kemungkinan pertukaran *medoid* dengan objek *non-medoid*, dan mengulangi proses sampai total biaya jarak tidak lagi menurun. Pada tahap interpretasi, objek yang menjadi medoid dapat dijadikan contoh transaksi atau pelanggan representatif dari setiap segmen. Inilah keunggulan utama *K-Medoids* dibandingkan *K-Means* dalam konteks bisnis, karena manajemen dapat melihat wakil data aktual dari setiap segmen, bukan hanya nilai rata-rata abstrak.

Implementasi teknis *K-Medoids* dalam penelitian ini menggunakan prosedur deterministik dan random restart. Inisialisasi pertama dibantu oleh hasil *K-Means* untuk memilih objek aktual terdekat dari pusat kelompok, kemudian ditambah dengan inisialisasi *farthest-first* dan *random restart*. Pada setiap iterasi, seluruh objek dialokasikan ke medoid terdekat menggunakan jarak *Euclidean*, kemudian medoid setiap klaster diperbarui dengan memilih objek yang memiliki total jarak terkecil terhadap anggota klaster tersebut. Proses berhenti ketika *medoid* tidak berubah atau jumlah iterasi maksimum tercapai. Skema ini tetap merepresentasikan prinsip dasar *K-Medoids* karena pusat klaster selalu berupa objek aktual, bukan rata-rata fitur [10], [13], [14].

Keputusan akhir algoritma terbaik tidak hanya ditentukan oleh angka *Silhouette Score*, tetapi juga oleh kemudahan interpretasi segmen. Dalam penelitian terapan, model yang memberikan skor sedikit lebih tinggi tetapi sulit diterjemahkan ke dalam kebijakan bisnis belum tentu paling berguna. Oleh sebab itu, penelitian ini menilai kualitas teknis dan kelayakan interpretasi secara bersamaan.

## 3. HASIL DAN ANALISIS

Tahap awal hasil penelitian menunjukkan bahwa dataset yang digunakan telah melalui pembersihan dan transformasi sehingga siap untuk pemodelan *clustering*. Data akhir yang dianalisis berjumlah 923 objek transaksi valid. Seluruh objek dianalisis berdasarkan dua atribut utama, yaitu *Quantity* dan Expected Revenue. Nilai *Expected Revenue* yang sebelumnya berbentuk teks mata uang berhasil dikonversi menjadi numerik, sedangkan data yang tidak lengkap, bernilai nol, atau merupakan baris ringkasan tidak digunakan dalam proses pemodelan. Setelah itu, data dinormalisasi melalui log transformation dan StandardScaler sehingga skala antaratribut menjadi sebanding.

Penentuan jumlah klaster pada *K-Means* dilakukan menggunakan Elbow Method. Hasil pengujian nilai k menunjukkan adanya penurunan SSE yang sangat tajam dari k = 1 ke k = 2, kemudian penurunan masih terlihat dari k = 2 ke k = 3. Setelah k = 3, penurunan SSE mulai melandai dan tidak menunjukkan pengurangan error yang signifikan. Pola tersebut mengindikasikan bahwa k = 3 merupakan titik siku yang paling representatif. Secara substantif, tiga klaster juga masuk akal untuk kebutuhan segmentasi pasar karena dapat diterjemahkan menjadi segmen bernilai rendah, menengah, dan tinggi. Penentuan k yang terlalu besar berisiko membuat segmentasi menjadi terlalu rinci dan sulit digunakan oleh manajemen, sedangkan k yang terlalu kecil dapat menghilangkan variasi penting dalam pola transaksi.

Penerapan *K-Means* dengan k = 3 menghasilkan tiga kelompok utama. Kelompok pertama merepresentasikan transaksi dengan *Quantity* rendah dan *Expected Revenue* rendah. Kelompok ini diberi label *Low Value* karena kontribusi volume dan nilai moneternya relatif terbatas. Kelompok kedua merepresentasikan transaksi dengan nilai menengah; sebagian pelanggan memiliki kuantitas pembelian yang tidak terlalu tinggi, tetapi nilai moneternya berada di atas kelompok rendah. Kelompok ini diberi label *Mid Value*. Kelompok ketiga merepresentasikan pelanggan dengan *Quantity* lebih tinggi dan *Expected Revenue* tinggi. Kelompok ini diberi label *High Value* karena berkontribusi besar terhadap pendapatan perusahaan. Hasil visualisasi scatter plot pada ruang data terstandardisasi menunjukkan bahwa *K-Means* mampu membentuk pemisahan yang cukup jelas, meskipun beberapa objek pada area transisi masih berdekatan dengan klaster lain.

Penentuan jumlah klaster pada *Agglomerative Hierarchical Clustering* dilakukan melalui dendrogram dengan Ward Linkage. Dendrogram memperlihatkan struktur penggabungan objek dari tingkat paling rendah hingga tingkat paling tinggi. Pada tingkat dissimilarity tertentu, terlihat tiga kelompok yang dapat dipertahankan sebagai struktur segmentasi. Hasil ini konsisten dengan *Elbow Method* pada *K-Means*. Konsistensi tersebut menunjukkan bahwa data penjualan PT XYZ memiliki kecenderungan pembentukan tiga kelompok alami pada dimensi *Quantity* dan Expected Revenue. *Hierarchical Clustering* kemudian dijalankan dengan jumlah klaster sebanyak tiga agar hasilnya dapat dibandingkan secara langsung dengan *K-Means*.

Hasil *Hierarchical Clustering* juga menghasilkan struktur segmen *Low Value*, *Mid Value*, dan *High Value*. Pola umumnya relatif sejalan dengan *K-Means*, yaitu terdapat kelompok transaksi kecil, kelompok transaksi menengah, dan kelompok transaksi bernilai tinggi. Perbedaan utama terletak pada cara algoritma membentuk batas klaster. *K-Means* membagi data berdasarkan kedekatan terhadap *centroid*, sedangkan *Hierarchical* membentuk kelompok berdasarkan proses penggabungan bertahap. Karena *Ward Linkage* berfokus pada minimisasi varians dalam klaster, beberapa objek pada wilayah transisi dapat ditempatkan berbeda dibandingkan hasil *K-Means*. Hal tersebut wajar karena kedua algoritma memiliki mekanisme optimasi yang berbeda.

Penerapan *K-Medoids* sebagai algoritma pembanding ketiga dilakukan dengan prinsip jumlah klaster yang sama, yaitu tiga klaster. Berbeda dari *K-Means* yang menggunakan *centroid* sebagai pusat klaster, *K-Medoids* menggunakan objek aktual sebagai medoid. Hasil *K-Medoids* juga menghasilkan tiga kelompok utama yang dapat diinterpretasikan sebagai *Low Value*, *Mid Value*, dan *High Value*. Pada model ini, segmen *Low Value* berisi 188 objek, *Mid Value* berisi 469 objek, dan *High Value* berisi 266 objek. Komposisi tersebut menunjukkan bahwa sebagian besar transaksi berada pada segmen menengah, sedangkan segmen tinggi memiliki proporsi cukup besar dan perlu menjadi fokus strategi retensi pelanggan.

Nilai *Silhouette Score K-Medoids* telah dihitung ulang dari dataset mentah yang sama, bukan diperkirakan dari skor *K-Means* atau *Hierarchical Clustering*. Hal ini penting karena setiap algoritma memiliki mekanisme pembentukan klaster yang berbeda. *K-Means* membentuk pusat klaster melalui *centroid*, *Hierarchical Clustering* membentuk struktur bertingkat melalui penggabungan klaster, sedangkan *K-Medoids* memilih objek aktual sebagai pusat klaster. Dengan menggunakan data, fitur, jumlah klaster, dan metrik evaluasi yang sama, perbandingan antarketiga algoritma menjadi lebih valid secara metodologis.

Evaluasi menggunakan *Silhouette Score* menunjukkan bahwa *K-Means* memperoleh nilai 0,4805301058, *Agglomerative Hierarchical Clustering* memperoleh nilai 0,4808089886, dan *K-Medoids* memperoleh nilai 0,4839631218. Dengan demikian, *K-Medoids* menjadi algoritma dengan skor tertinggi pada dataset ini. Selisih *K-Medoids* terhadap *K-Means* adalah sekitar 0,0034330160, sedangkan selisih *K-Medoids* terhadap *Hierarchical Clustering* adalah sekitar 0,0031541331. Meskipun perbedaannya tidak besar, hasil ini menunjukkan bahwa pendekatan berbasis medoid sedikit lebih mampu membentuk klaster yang kohesif dan terpisah pada data penjualan yang telah ditransformasi.

Tabel 3. Perbandingan nilai Silhouette Score

| Algoritma | Jumlah Klaster | Silhouette Score | Interpretasi |
|---|---|---|---|
| *K-Means* | 3 | 0,4805301058 | Baseline partitional *clustering* berbasis *centroid*; hasil baik, tetapi bukan skor tertinggi pada pengujian ulang. |
| *Agglomerative Hierarchical Clustering* | 3 | 0,4808089886 | Sedikit lebih tinggi daripada *K-Means*; berguna untuk validasi struktur bertingkat melalui dendrogram. |
| *K-Medoids* | 3 | 0,4839631218 | Skor tertinggi; dipilih sebagai model akhir karena pusat klaster berupa objek aktual atau medoid. |

Dengan ditambahkannya *K-Medoids*, interpretasi perbandingan algoritma menjadi lebih komprehensif. *K-Means* unggul dari sisi kesederhanaan dan efisiensi, *Hierarchical Clustering* unggul dari sisi eksplorasi struktur bertingkat, sedangkan *K-Medoids* unggul dari sisi representasi pusat klaster yang berbasis objek aktual. Pada data penjualan yang memiliki nilai transaksi sangat beragam, *K-Medoids* memberikan hasil evaluasi tertinggi sehingga dapat dipertimbangkan sebagai model akhir. Namun, karena selisih *Silhouette Score* antarketiga algoritma relatif kecil, keputusan model akhir tetap perlu mempertimbangkan aspek teknis, stabilitas implementasi, dan kebutuhan interpretasi bisnis.

Keunggulan *K-Medoids* dalam penelitian ini dapat dijelaskan dari karakteristik data penjualan yang memiliki nilai transaksi sangat bervariasi. Pada kondisi tersebut, pusat klaster berbasis *centroid* seperti pada *K-Means* dapat dipengaruhi oleh nilai ekstrem, sedangkan medoid selalu merupakan objek aktual yang lebih mudah

dihubungkan dengan profil transaksi nyata. *Hierarchical Clustering* tetap memberikan nilai analitis melalui dendrogram yang memperlihatkan struktur penggabungan objek, sedangkan *K-Means* tetap relevan sebagai model pembanding yang cepat dan sederhana. Dengan demikian, *K-Medoids* dipilih sebagai model akhir secara numerik, *K-Means* berperan sebagai baseline partitional *clustering*, dan *Hierarchical Clustering* berperan sebagai validasi struktur bertingkat [6], [9], [10], [11].

Tabel 4. Perbandingan karakteristik *K-Means*, *Hierarchical Clustering*, dan *K-Medoids*

| Aspek | *K-Means* | *Hierarchical Clustering* | *K-Medoids* |
|---|---|---|---|
| Tipe algoritma | Partitional berbasis *centroid* | Hierarchical/agglomerative berbasis penggabungan bertahap | Partitional berbasis medoid |
| Pusat klaster | *Centroid*/rata-rata fitur | Tidak memiliki pusat tunggal; struktur dilihat melalui dendrogram | Medoid atau objek aktual dalam data |
| Kelebihan utama | Sederhana, cepat, mudah divisualisasikan | Menjelaskan struktur hubungan bertingkat antardata | Lebih mudah diinterpretasikan dan relatif lebih tahan terhadap outlier |
| Keterbatasan | Sensitif terhadap *centroid* awal dan outlier | Komputasi lebih berat untuk data besar | Komputasi dapat lebih berat dibanding *K-Means* |
| Peran dalam penelitian | Baseline pembanding berbasis *centroid* yang cepat dan mudah diterapkan. | Pembanding hierarkis untuk membaca struktur penggabungan dan validasi dendrogram. | Model akhir terbaik berdasarkan *Silhouette Score* dan interpretasi medoid aktual. |

Tabel tersebut memperlihatkan bahwa penambahan *K-Medoids* tidak hanya memperluas ruang pembandingan algoritma, tetapi juga memperkuat analisis metodologis. Jika tujuan utama perusahaan adalah model yang cepat dan mudah diterapkan, *K-Means* tetap kuat. Jika tujuan utama adalah memahami struktur bertingkat antardata, *Hierarchical Clustering* lebih informatif. Jika tujuan utama adalah memperoleh wakil segmen yang benar-benar berasal dari data aktual, *K-Medoids* menjadi pilihan yang relevan. Dalam hasil pengujian ulang pada dataset mentah, *K-Medoids* memperoleh skor tertinggi sehingga lebih layak dipilih sebagai model akhir untuk interpretasi segmentasi pasar.

Sebagai tambahan, *K-Medoids* menghasilkan pusat klaster berupa objek aktual. Pada hasil pemodelan, medoid *Low Value* berada pada transaksi dengan *Quantity* 1 dan *Expected Revenue* Rp7.150.000, medoid *Mid Value* berada pada transaksi dengan *Quantity* 1 dan *Expected Revenue* Rp78.000.000, sedangkan medoid *High Value* berada pada transaksi dengan *Quantity* 7 dan *Expected Revenue* Rp455.000.000. Informasi ini memperkuat interpretasi bisnis karena setiap pusat segmen dapat dijelaskan menggunakan profil transaksi nyata tanpa harus membuka identitas pelanggan.

Hasil distribusi pelanggan berdasarkan model *K-Medoids* menunjukkan bahwa segmen *Mid Value* merupakan kelompok terbesar dengan 469 objek atau sekitar 50,81% dari total data. Segmen *High Value* berjumlah 266 objek atau sekitar 28,82%, sedangkan segmen *Low Value* berjumlah 188 objek atau sekitar 20,37%. Komposisi ini memberikan gambaran bahwa pasar PT XYZ didominasi oleh transaksi bernilai menengah, tetapi terdapat kelompok bernilai tinggi yang cukup besar. Kelompok *High Value* perlu mendapat prioritas karena kontribusi moneternya lebih besar, sedangkan kelompok *Mid Value* berpotensi dikembangkan melalui strategi *upselling*, *cross-selling*, dan *bundling* produk.

Tabel 5. Distribusi jumlah pelanggan per segmen

| Segmen | Jumlah Pelanggan | Persentase | Makna Bisnis |
|---|---|---|---|
| *Low Value* | 188 | 20,37% | *Quantity* dan *Expected Revenue* relatif rendah; dapat dikelola dengan strategi pemasaran efisien. |
| *Mid Value* | 469 | 50,81% | Kelompok terbesar; potensial untuk *upselling, cross-selling,* dan *bundling* produk. |
| *High Value* | 266 | 28,82% | Kontribusi moneter tinggi; perlu prioritas retensi dan layanan khusus. |

Analisis sektor dilakukan secara terbatas karena dataset tidak menyediakan atribut sektor secara eksplisit. Oleh karena itu, sektor privat dan pemerintah/negeri diidentifikasi melalui inferensi kata kunci pada *Opportunity Name*, misalnya istilah dinas, kementerian, pemerintah daerah, DPRD, sekolah negeri, atau lembaga negara. Berdasarkan inferensi tersebut, sektor privat masih mendominasi pada seluruh segmen. Pada segmen *Low Value*, sektor privat mencapai sekitar 78,2% dan pemerintah/negeri 21,8%. Pada segmen *Mid Value*, sektor privat sekitar 72,1% dan pemerintah/negeri 27,9%. Pada segmen *High Value*, sektor privat sekitar 71,1% dan

pemerintah/negeri 28,9%. Hasil ini perlu divalidasi kembali oleh perusahaan jika ingin digunakan untuk keputusan strategis berbasis sektor.

Tabel 6. Distribusi sektor pada setiap segmen

| Segmen | Sektor Privat | Sektor Pemerintah/Negeri | Interpretasi |
|---|---|---|---|
| *Low Value* | 78,2% | 21,8% | Berdasarkan inferensi kata kunci, transaksi rendah lebih banyak berasal dari sektor privat. |
| *Mid Value* | 72,1% | 27,9% | Sektor privat masih dominan; segmen ini menjadi basis utama potensi upselling. |
| *High Value* | 71,1% | 28,9% | Sektor privat tetap dominan, tetapi pemerintah/negeri berkontribusi pada transaksi bernilai tinggi. |

Jika dikaitkan dengan penelitian sebelumnya, hasil ini mendukung pandangan bahwa *clustering* dapat membantu perusahaan memahami pola penjualan dan preferensi pelanggan. Studi data mining pada perusahaan menunjukkan bahwa teknik *clustering* mampu mengungkap pola yang tidak terlihat melalui rekapitulasi sederhana [21]. Penelitian segmentasi pasar menggunakan *K-Means* juga menunjukkan bahwa pengelompokan pelanggan dapat membantu penyusunan strategi pemasaran yang lebih terarah [6]. Sementara itu, penelitian yang membandingkan *K-Means* dan *Hierarchical Clustering* menegaskan pentingnya evaluasi menggunakan indeks validitas seperti *Silhouette Score* [11]. Penambahan *K-Medoids* dalam penelitian ini memperkaya perbandingan karena memberikan perspektif pusat klaster berbasis objek aktual [10], [13], [14].

Namun demikian, kualitas *Silhouette Score* yang berada di sekitar 0,48 perlu dibaca secara hati-hati. Nilai tersebut tidak dapat diartikan sebagai kegagalan model, tetapi menunjukkan bahwa karakteristik pelanggan pada data penjualan memiliki struktur klaster sedang dan masih terdapat area tumpang tindih antarsegmen. Hal ini wajar pada data penjualan bisnis karena pelanggan tidak selalu terbagi secara ekstrem ke dalam kelompok rendah, menengah, dan tinggi. Beberapa pelanggan dapat berada pada batas antara *Mid Value* dan *High Value*, terutama ketika *Quantity* tidak terlalu besar tetapi *Expected Revenue* tinggi, atau sebaliknya.

Implikasi praktis penelitian ini dapat dirumuskan ke dalam tiga strategi. Pertama, perusahaan perlu menjaga pelanggan *High Value* melalui pelayanan prioritas, komunikasi proaktif, penawaran produk pelengkap, dan pengelolaan relasi jangka panjang. Kedua, perusahaan perlu mengembangkan segmen *Mid Value* karena jumlahnya paling besar dan memiliki potensi peningkatan nilai transaksi. Program bundling, diskon berbasis volume, dan rekomendasi produk berdasarkan kebutuhan pelanggan dapat diarahkan pada kelompok ini. Ketiga, segmen *Low Value* tetap perlu dikelola secara efisien melalui pendekatan digital marketing, katalog produk, atau kampanye penawaran standar agar biaya akuisisi dan pelayanan tidak melebihi potensi kontribusinya. Strategi tersebut menunjukkan bahwa hasil *clustering* dapat menjadi dasar diferensiasi layanan pemasaran.

Tabel 7. Rekomendasi strategi pemasaran berdasarkan hasil *clustering*

| Segmen | Rekomendasi Strategi | Tujuan Manajerial |
|---|---|---|
| *Low Value* | Efisiensi biaya pemasaran, kampanye digital standar, dan edukasi produk dasar. | Menjaga peluang transaksi tanpa biaya layanan berlebihan. |
| *Mid Value* | *Upselling*, *cross-selling*, *bundling* produk, dan penawaran berbasis kebutuhan. | Mendorong sebagian pelanggan berpindah ke segmen *High Value*. |
| *High Value* | Retensi pelanggan prioritas, account management, layanan khusus, dan pengelolaan relasi jangka panjang. | Melindungi kontribusi pendapatan utama perusahaan. |

Dari sisi akademik, penelitian ini menunjukkan bahwa perbandingan algoritma perlu dilakukan dengan prosedur yang konsisten. Ketiga algoritma dijalankan pada data yang sama, fitur yang sama, jumlah klaster yang setara, dan skema evaluasi yang sama. Tanpa konsistensi tersebut, perbandingan dapat menghasilkan kesimpulan yang bias. Hasil revisi ini memperlihatkan bahwa *K-Medoids* memperoleh *Silhouette Score* tertinggi, tetapi selisihnya kecil dibandingkan *K-Means* dan *Hierarchical Clustering*. Hal ini menguatkan bahwa interpretasi akhir tidak boleh hanya bergantung pada satu angka evaluasi, melainkan perlu dikaitkan dengan konteks bisnis dan tujuan penggunaan model.

Secara keseluruhan, hasil penelitian memperlihatkan bahwa *K-Means*, *Hierarchical Clustering*, dan *K-Medoids* sama-sama dapat digunakan untuk segmentasi pasar berdasarkan *Quantity* dan Expected Revenue, sepanjang seluruh algoritma dijalankan melalui tahapan preprocessing yang benar. Pada dataset ini, *K-Medoids* memberikan nilai *Silhouette Score* tertinggi sebesar 0,4839631218, diikuti *Hierarchical Clustering* sebesar 0,4808089886, dan *K-Means* sebesar 0,4805301058. Oleh karena itu, *K-Medoids* dipilih sebagai model akhir,

sedangkan *K-Means* dan *Hierarchical Clustering* digunakan sebagai pembanding yang memperkuat validitas analisis.

Keterbatasan utama hasil ini adalah jumlah atribut yang digunakan masih terbatas pada *Quantity* dan Expected Revenue. Dua atribut tersebut memang mampu mewakili volume dan nilai transaksi, tetapi belum sepenuhnya menggambarkan perilaku pelanggan. Segmentasi pasar akan menjadi lebih kaya apabila penelitian berikutnya menambahkan frekuensi pembelian, recency transaksi, margin keuntungan, kategori produk, wilayah pelanggan, jenis institusi, dan sektor pelanggan yang telah tervalidasi. Penambahan atribut tersebut berpotensi meningkatkan kualitas *Silhouette Score* dan memperjelas batas antarklaster. Selain itu, penelitian berikutnya dapat membandingkan metode validasi lain seperti Davies-Bouldin Index dan Calinski-Harabasz Index agar pemilihan model akhir tidak hanya bergantung pada satu metrik evaluasi.

Dari perspektif manajerial, penggunaan *K-Medoids* sebagai model akhir juga memberikan implikasi yang lebih praktis. Setiap segmen dapat dijelaskan melalui medoid sebagai contoh transaksi representatif, sehingga tim pemasaran dapat memahami karakteristik pusat dari kelompok *Low Value*, *Mid Value*, dan *High Value*. Segmen *Low Value* dapat diarahkan pada strategi efisiensi biaya dan kampanye digital massal, segmen *Mid Value* dapat diarahkan pada strategi peningkatan nilai transaksi melalui *bundling* dan cross-selling, sedangkan segmen *High Value* perlu diprioritaskan melalui account management, komunikasi personal, dan layanan purna jual yang lebih intensif. Dengan cara ini, hasil data mining tidak berhenti pada evaluasi teknis, tetapi diterjemahkan menjadi rekomendasi bisnis yang dapat dijalankan.

Walaupun *K-Medoids* memperoleh skor tertinggi, selisih nilai antaralgoritma masih relatif kecil. Kondisi ini mengindikasikan bahwa ketiga algoritma membaca struktur data yang hampir serupa, yaitu adanya kelompok transaksi rendah, menengah, dan tinggi. Dengan demikian, hasil penelitian tidak boleh ditafsirkan bahwa *K-Means* atau *Hierarchical Clustering* tidak layak digunakan. *K-Means* tetap relevan jika perusahaan membutuhkan model yang sederhana, cepat, dan mudah diimplementasikan dalam sistem dashboard penjualan. *Hierarchical Clustering* tetap relevan untuk tahap eksplorasi karena dendrogram membantu memahami kedekatan antardata dan kemungkinan struktur klaster alternatif. Namun, jika keputusan akhir diarahkan pada model yang paling baik berdasarkan *Silhouette Score* serta memiliki pusat klaster berupa objek aktual, maka *K-Medoids* menjadi pilihan yang lebih kuat pada dataset ini.

Validasi ulang *K-Medoids* pada dataset mentah memberikan kontribusi penting terhadap kualitas naskah karena sebelumnya *K-Medoids* hanya diposisikan sebagai algoritma pembanding konseptual. Setelah data mentah dihitung ulang, nilai *Silhouette Score K-Medoids* terbukti lebih tinggi dibandingkan *K-Means* dan Agglomerative *Hierarchical Clustering*. Secara metodologis, hal ini menunjukkan bahwa pemilihan pusat klaster berbasis medoid lebih sesuai untuk data penjualan PT XYZ yang memiliki variasi nilai transaksi cukup besar. Pada data dengan distribusi tidak simetris, *centroid* dapat bergeser karena pengaruh transaksi bernilai sangat tinggi, sedangkan medoid tetap berada pada titik data aktual yang paling representatif terhadap anggota klasternya. Oleh karena itu, *K-Medoids* tidak hanya memberikan angka evaluasi terbaik, tetapi juga memberikan dasar interpretasi yang lebih mudah dijelaskan kepada manajemen karena pusat segmen dapat dikaitkan dengan profil transaksi nyata.

## 4. KESIMPULAN

Penelitian ini menyimpulkan bahwa segmentasi pasar berbasis data penjualan PT XYZ dapat dibentuk menggunakan *K-Means*, Agglomerative *Hierarchical Clustering*, dan *K-Medoids* dengan atribut utama *Quantity* dan Expected Revenue. Setelah dilakukan pembersihan, transformasi logaritmik, dan standardisasi, diperoleh 923 data valid untuk pemodelan. Hasil evaluasi menggunakan *Silhouette Score* menunjukkan bahwa *K-Medoids* memperoleh nilai tertinggi sebesar 0,4839631218, disusul *Agglomerative Hierarchical Clustering* sebesar 0,4808089886, dan *K-Means* sebesar 0,4805301058. Dengan demikian, *K-Medoids* dipilih sebagai model akhir karena memberikan kualitas klaster terbaik secara numerik dan memiliki keunggulan interpretatif berupa medoid sebagai objek aktual. Hasil segmentasi membentuk tiga kelompok, yaitu *Low Value* sebanyak 188 objek, *Mid Value* sebanyak 469 objek, dan *High Value* sebanyak 266 objek. Secara praktis, perusahaan dapat menggunakan hasil ini untuk menyusun strategi pemasaran yang berbeda pada setiap segmen, terutama dengan mendorong pengembangan pelanggan *Mid Value* dan mempertahankan pelanggan *High Value*. Penelitian selanjutnya disarankan menambahkan variabel lain seperti frekuensi transaksi, jenis produk, margin keuntungan, wilayah pelanggan, dan sektor pelanggan yang tervalidasi agar kualitas segmentasi dapat meningkat.

## UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada PT XYZ, khususnya bagian marketing yang telah menyediakan akses data penjualan untuk kebutuhan penelitian. Ucapan terima kasih juga disampaikan kepada rekan-rekan dosen di Program Studi Sistem Informasi, Universitas Pembangunan Jaya yang telah memberikan masukan dalam penyusunan artikel ini. Seluruh data yang digunakan dalam penelitian disajikan secara agregat untuk menjaga kerahasiaan informasi perusahaan dan pelanggan.

**REFERENSI**

[1] S. Gea, "Pengaruh Segmentasi Pasar Terhadap Peningkatan Volume Penjualan," *J. Akuntansi, Manaj. dan Ekon.*, vol. 1, no. 1, pp. 48–54, 2022, doi: 10.56248/jamane.v1i1.12.

[2] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma *K-Means Clustering* Dalam Evaluasi Hasil Pembelajaran Siswa," *J. Ilm. Inform. dan Ilmu Komput.*, vol. 3, no. 1, pp. 46–56, 2024, doi: 10.58602/jima-ilkom.v3i1.26.

[3] M. Norshahlan, H. Jaya, and R. Kustini, "Penerapan Metode *Clustering* Dengan Algoritma *K-Means* Pada Pengelompokan Data Calon Siswa Baru," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 2, no. 6, p. 1042, 2023, doi: 10.53513/jursi.v2i6.9148.

[4] E. Elni Arbaeti, A. M. Hara Pardede, and L. A. Nur Kadim, "Application of *K-Means Clustering* Algorithm To Analyze Insurance Company Business (Case Study: Pt. Jasindo Insurance)," *J. Math. Technol.*, vol. 2, no. 2, pp. 173–192, 2023, doi: 10.63893/matech.v2i2.161.

[5] V. Alvianatinova, I. Ali, N. Rahaningsih, and A. Bahtiar, "Penerapan Algoritma *K-Means Clustering* Dalam Pengelompokan Data Penjualan Supermarket Berdasarkan Cabang (Branch)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1529–1535, 2024, doi: 10.36040/jati.v8i2.8993.

[6] R. F. M. Bayu Wibawa, Mahendar Dwi Payana, "*Clustering* Data Pelanggan Menggunakan *K-Means* untuk Segmentasi Pasar pada Sentra UMKM di Kota Banda Aceh," *JPKMI - J. Pengabdi. Kpd. Masy. Bid. Inotec*, vol. 15, no. 1, pp. 72–86, 2024, doi: 10.25130/sc.24.1.6.

[7] O. Eric U. and O. Michael O., "Overview of *Agglomerative Hierarchical Clustering* Methods," *Br. J. Comput. Netw. Inf. Technol.*, vol. 7, no. 2, pp. 14–23, 2024, doi: 10.52589/bjcnit-cv9poogw.

[8] R. P. Justitia, N. Hidayat, and E. Santoso, "Implementasi Metode *Agglomerative Hierarchical Clustering* Pada Segmentasi Pelanggan Barbershop (Studi Kasus : RichDjoe Barbershop Malang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 3, pp. 1048–1054, 2021, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/8730

[9] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*.

[10] A. A. D. Sulistyawati and M. Sadikin, "Penerapan Algoritma *K-Medoids* Untuk Menentukan Segmentasi Pelanggan," *Sist. J. Sist. Inf.*, vol. 10, no. 3, p. 516, 2021, doi: 10.32520/stmsi.v10i3.1332.

[11] S. D. K. Wardani, A. S. Ariyanto, M. Umroh, and D. Rolliawati, "Comparison of *K-Means*, Db Scanner & *Hierarchical Clustering* Method Results for Market Segmentation Analysis," *JIKO (Jurnal Inform. dan Komputer)*, vol. 7, no. 2, p. 191, 2023.

[12] F. D. Wahyuningtyas, A. Arafat, A. Stiawan, and D. Rolliawati, "Komparasi Algoritma Hierarchical, *K-Means*, dan DBSCAN pada Analisis Data Penjualan Melalui Facebook," *Explor. J. Sist. Inf. dan Telemat.*, vol. 14, no. 1, p. 7, 2023, doi: 10.36448/jsit.v14i1.2931.

[13] M. D. Salman, N. R. Pratama, and M. N. F. A, "Comparison of *K-Means* and *K-Medoids Clustering* Algorithm Performance in Grouping Schools in Riau Province Based on Availability of Facilities and Infrastructure Perbandingan Kinerja Algoritma *Clustering K-Means* dan *K-Medoids* dalam Pengelompokan Sekolah di," *MALCOMIndonesian J. Mach. Learn. Comput. Sci.*, vol. 5, no. July, pp. 797–806, 2025.

[14] H. M. Ilmi, M. Kurniawan, U. Al Faruq, and R. R. Muhima, "Comparison of *K-Means* and *K-Medoids* for Hotspot Data *Clustering* on the Island of Kalimantan," *J. SimanteC*, vol. 13, no. 1, pp. 33–40, 2024.

[15] M. A. Hasanah, S. Soim, and A. S. Handayani, "M. A. Hasanah, S. Soim, dan A. S. Handayani, 'Implementasi CRISP-DM model menggunakan metode decision tree dengan algoritma CART untuk prediksi curah hujan berpotensi banjir,' 2021.," *J. Appl. Informatics Comput.*, vol. 5, no. 2, p. 103, 2021, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[16] R. A. Azizah, F. Bachtiar, and S. Adinugroho, "Klasifikasi Kinerja Akademik Siswa Menggunakan Neighbor Weighted K-Nearest Neighbor dengan Seleksi Fitur Information Gain," *J. Teknol. Inf. dan*

*Ilmu Komput.*, vol. 9, no. 3, pp. 605–614, 2022, doi: 10.25126/jtiik.2022935751.

[17] R. Ishak, Nurmawanti, and Amiruddin, "Optimasi *K-Means* pada *Clustering* Penyakit Ibu Hamil Menggunakan Random Forest," *Jambura J. Electr. Electron. Eng.*, vol. 7, no. 1, p. 41, 2024.

[18] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada *K-Means Clustering*," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 14, no. 2, pp. 277–288, 2023, doi: 10.24176/simet.v14i2.9630.

[19] A. M. A. S. Sidebang, "Pemetaan Kemiskinan Digital Kabupaten/Kota di Sumatera Utara Menggunakan Ward *Hierarchical Clustering* Digital Poverty Mapping of Regencies/Cities in North Sumatra Using Ward *Hierarchical Clustering*," 2024.

[20] P. Sinulingga, N. I. Siregar, and A. H. Lubis, "Klasterisasi Lokasi Wisata di Indonesia dengan Menggunakan Algoritma *Hierarchical Clustering*," *J. Comput. Informatics Res.*, vol. 4, no. 3, p. 361, 2025, doi: 10.47065/comforch.v4i3.2079.

[21] M. Risqi Ananda, N. Sandra, E. Fadhila, A. Rahma, and N. Nurbaiti, "Data Mining dalam Perusahaan PT Indofood Lubuk Pakam," *Com. Commun. Inf. Technol. J.*, vol. 2, no. 1, pp. 108–119, 2023, doi: 10.47467/comit.v2i1.124.