p-ISSN: 2502-3691
e-ISSN: 2502-4361

KMeans Clustering Menggunakan RapidMiner dalam Segmentasi Pelanggan dengan Evaluasi Davies Bouldin Index Untuk Menentukan Jumlah Cluster Paling Optimal

Pingky Septiana Ananda[1*)], Eko Sediono[2], Irwan Sembiring[3]
[1,2,3]Magister Sistem Informasi, Fakultas Teknik Informasi, Universitas Kristen Satya Wacana
[1,2,3] Jl. Jl. O. Notohamidjojo, No. 1-10 Salatiga
email: [1]972021013@student.uksw.edu, [2]eko@uksw.edu, [3]irwan@uksw.edu

***Abstract* – A small and medium business (UKM) is driven to provide top-notch customer service by fierce competition in the business sector. Customer segmentation is a concept that refers to combining potential customers into certain groups or segments where through this customer segmentation, sellers can reach the right customers This study aims to select the most ideal number of groups in customer segmentation. research results obtained by using the K-Means solving procedure in clustering can provide customer segmentation results that are in line with expectations, so that sellers can easily understand the characteristics of their customers based on their clusters, income, expenditure as well as age and gender. Then with the evaluation Davies Boildin also provides a solution for selecting the right number of clusters so that the performance is more optimal and produces more accurate customer segmentation results.**

***Keywords:* K-Means Clustering, Data Mining, Daviel Bouldin Index**

***Abstrak* – Usaha kecil dan menengah (UKM) didorong untuk memberikan layanan pelanggan terbaik oleh persaingan ketat di sektor bisnis. Segmentasi pelanggan merupakan suatu konsep yang mengacu pada penggabungan calon pelanggan ke dalam kelompok atau segmen tertentu dimana melalui segmentasi pelanggan ini, penjual bisa menjangkau pelanggan yang tepat. Penelitian ini bertujuan untuk memilih jumlah grup yang paling ideal dalam segmentasi pelanggan. Hasil penelitian yang diperoleh dengan memakai prosedur pemecahan K-Means dalam clustering data pelanggan dapat memberikan hasil segmentasi pelanggan yang sesuai dengan harapan, sehingga penjual dapat dengan mudah memahami karakteristik pelanggannya berdasarkan clusternya, pemasukannya, pengeluarannya maupun umur dan gendernya. Kemudian dengan evaluasi Davies Boildin juga memberikan solusi pemilihan jumlah cluster yang tepat sehingga performanya lebih optimal dan menghasilkan hasil segmentasi pelanggan yang lebih akurat.**

***Kata Kunci* – K-Means Clustering, Data Mining, Daviel Bouldin Index**

I. PENDAHULUAN

Segmentasi pelanggan merupakan suatu konsep yang mengacu pada penggabungan calon pelanggan ke dalam kelompok atau segmen tertentu dimana melalui segmentasi pelanggan ini, penjual bisa menjangkau pelanggan yang tepat. Memahami segmentasi pasar adalah bagian penting untuk membuat strategi yang sesuai[1].

Segmentasi pelanggan berguna untuk menyusun strategi produk, penjualan, dan pemasaran yang sesuai dengan target pelanggannya. Segmentasi pelanggan adalah cara untuk memperkuat siklus pengembangan produk yang dapat mengurangi risiko kegagalan pemasaran dan memudahkan penjual untuk mempersonalisasi promosi pemasaran mereka ke pelanggan yang tepat.

Saat ini untuk melakukan analisa segmentasi pelanggan sangatlah banyak dengan berbagai kelebihannya masing-masing, namun saat ini penulis akan menggunakan prosedur pemecahan K-Means menggunakan evaluasi Davies Bouldin Index dalam menentukan jumlah kluster yang paling optimal dalam dataset pelanggan sehingga dengan jumlah cluster (kelompok) yang tepat maka tentu hasil segmentasinya juga menjadi lebih akurat[2].

Datamining artinya proses menemukan hubungan baru yg berguna, pola dan isu terkini dengan menambang gudang informasi yang tak terhitung jumlahnya, memanfaatkan kemajuan pengenalan desain seperti wawasan dan metode numerik [3].

Terdapat beberapa masalah yang harus dipertimbangkan saat menggunakan strategi K-Means menggabungkan model pengelompokan yang berbeda, memilih model yang paling tepat untuk dataset yang akan dibedah, ketidakmampuan untuk menggabungkan, pengenalan anomali, keadaan setiap kumpulan dan menutupi perselisihan..

Tujuan penelitian ini adalah untuk menentukan jumlah cluster yang paling optimal dalam segmentasi pelanggan.

II. PENELITIAN YANG TERKAIT

Hasil penelitian yang relevan digunakan untuk perbandingan atau sebagai acuan dalam penelitian ini. Segmentasi pelanggan adalah proses pembagian data-data pelanggan menjadi kelompok-kelompok yang dapat dianalisa lebih jauh lagi sesuai dengan karakteristik tertentu seperti umur, gender, pemasukannya, pengeluarannya, frekuensi pembelian dan lain-lain. Menurut Christina Deni Rumiarti,

p-ISSN: 2502-3691
e-ISSN: 2502-4361

Indra Budi, Pembagian klien dalam pandangan RFM membentuk dua kelompok ideal, khususnya klien periodik dan klien lesu. Pembagian klien sesuai jumlah jenis buku yang dibeli terdiri dari 3 kelompok ideal, yaitu rendah, sedang dan tinggi[4].

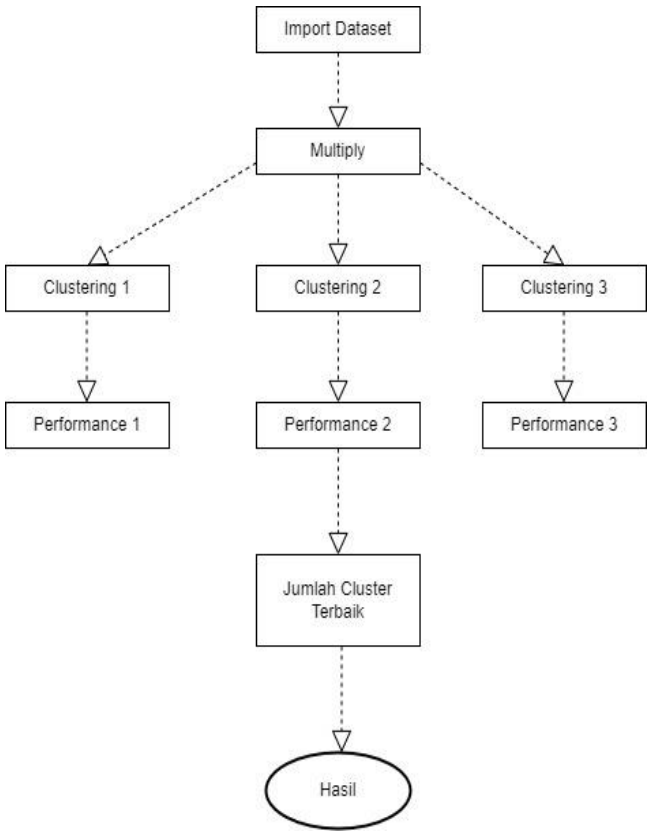
Penelitian lain yang diteliti adalah pemanfaatan teknik pengelompokan dengan metodologi K-implies settling untuk divisi klien pada CV. sinar surya. Penelitian tersebut menggunakan model untuk segmentasi RFM (Recency, Recurrence, and Money related), produk yang digunakan untuk menangani model RFM adalah Microsoft Succeed adaptation 2010 sedangkan untuk memasak teknik pengelompokan adalah Quick Excavator varian 9.0. awal pembagian ini membagi client menjadi 2 group [5].

Dari penelitian- penelitian yang sebelumnya Pemanfaatan kalkulasi K Medoids untuk menentukan Client Division, pengujian ini menerapkan prosedur pengelompokan dengan menggunakan kalkulasi K-Medoids untuk melakukan transaksi pertukaran dataset untuk menentukan client division[6]. Susunan strategi pemasaran dipengaruhi oleh jenis dan atribut klien di setiap kelompok atau fragmen klien yang dibingkai. Pengujian legitimasi kelompok menggunakan Outline Record dan Davies Bouldin File dilakukan untuk menentukan jumlah kelompok yang ideal.

III. METODE PENELITIAN

A. Kerangka Penelitian

Penelitian yang dilakukan menggunakan Metode / algoritma K-Means Clustering di RapidMiner dalam Segmentasi Pelanggan dengan Evaluasi Davies Bouldin Index untuk menentukan jumlah cluster paling optimal.



B. Data Understanding

Dalam pembuatan klasterisasi segmentasi pelanggan memerlukan data acuan berupa data pelanggan dari sebuah pusat perbelanjaan yang diambil dari situs kaggle.com. Data yang diambil dan diolah oleh penulis berisi data gender, usia, pendapatan dan pengeluaran pelanggan. Dalam pengolahannya penulis menentukan presentase angka pemasukan dan pengeluaran berdasarkan usia maupun gendernya[7].

C. RapidMiner

RapidMiner adalah pemrograman yang bersifat terbuka *(open source)*. RapidMiner adalah jawaban untuk membedah penambangan terukur, penambangan teks, dan pemeriksaan pendahuluan. RapidMiner menggunakan berbagai metode yang mencerahkan dan cerdas untuk memberikan sedikit pengetahuan kepada klien sehingga mereka dapat memilih pilihan yang paling ideal[8].

RapidMiner memiliki sekitar 500 administrator penambangan informasi, termasuk administrator untuk input, hasil, prapemrosesan informasi, dan representasi.

RapidMiner adalah aplikasi independen untuk penyelidikan informasi serta mesin penambangan informasi yang dapat dimasukkan ke dalam itemnya sendiri. RapidMiner disusun menggunakan bahasa Java sehingga dapat menangani setiap kerangka kerja.

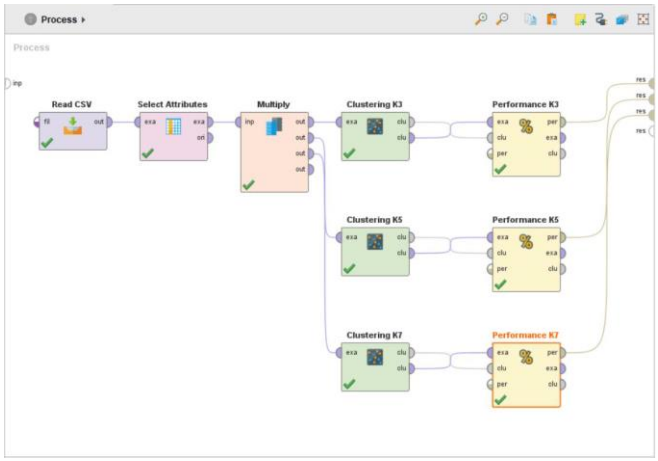
RapidMiner baru-baru ini disebut YALE (Satu lagi Iklim Pembelajaran), di mana terjemahan yang mendasarinya dibuat pada tahun 2001 oleh Ralf Klin kenberg, Ingo Mierswa, dan Simon Fischer di Unit Kesadaran Buatan Manusia dari College of Dortmund.

RapidMiner tersebar di bawah izin AGPL (Izin Populasi Keseluruhan Affero GNU) varian tiga. Hingga saat ini, banyak sekali aplikasi telah dibuat yang melibatkan RapidMiner di lebih dari 40 negara. RapidMiner menjadi perangkat lunak sumber terbuka untuk penambangan data, tidak diragukan lagi karena produk ini sudah terkenal di dunia.

p-ISSN: 2502-3691
e-ISSN: 2502-4361

IV. HASIL DAN PEMBAHASAN

Proses Clustering menggunakan algoritma K-Means di RapidMiner



Gambar 1. Proses Clustering

Keterangan :
1. Read CSV : Digunakan untuk mengimport dataset pelanggan
2. Select Attributes : Memilih atribut apa saja yang akan digunakan
3. Multiply : Untuk share dataset agar bisa digunakan oleh 3 K-means
4. K-means : Untuk clustering data
5. Performance : Untuk mengevaluasi kinerja clustering K-means

Penggunaan Algoritma K-Means

Algoritma ini yang penulis gunakan karena dihadapkan pada masalah yang penyelesaiannya membutuhkan proses segmentasi atau pengelompokkan pelanggan menjadi bebeberapa cluster, dalam hal ini penulis menggunakan 3 operator K-Means dengan jumlah cluster yang berbeda yaitu 3, 5 dan 7 dimana nanti dari ketiga operator K-Means ini akan dipilih salah satu yang paling optimal melalui evaluasi Davies Bouldin Index[9].

Hal ini dilakukan karena seperti yang sudah umum diketahui kalau salah satu kelemahan algoritma K-Means adalah sensitif terhada jumlah cluster yang artinya perbedaan jumlah cluster dapat berpengaruh pada hasil segmentasinya. Sebenarnya hal ini bisa ditangani dengan menggunakan Elbow Method, dengan menghitung dan membandingkan nilai WCSS atau juga bisa diperbaiki dengan teknik k-Medoids.

Penggunaan Davies Bouldin

Penulis menggunakan Davies Bouldin Index untuk mengevaluasi performance clustering K-Means karena di dalam Davies Bouldin ini sendiri terdapat dua indikator penting yang gunakan yaitu :

1. Indikator Homogenitas

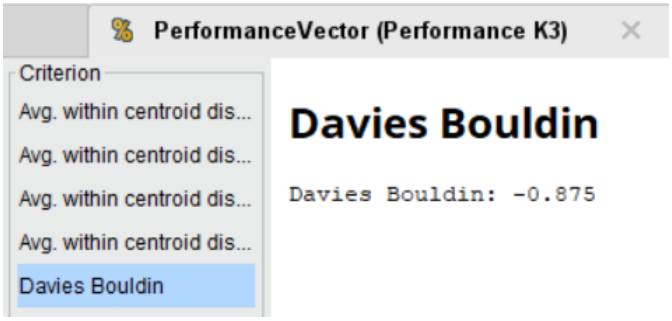
Dalam indikator ini akan dilihat seberapa mirip/kesamaan antara satu anggota dengan anggota yg lain dalam satu cluster. Makin kecil nilai Davies Bouldinnya berarti makin mirip antar anggotanya yang artinya semakin baik performance-nya, sebaliknya makin besar nilai Davies Bouldinnya maka makin tidak mirip antar anggota yang berarti kurang baik performance-nya[10].

2. Indikator Heterogenitas

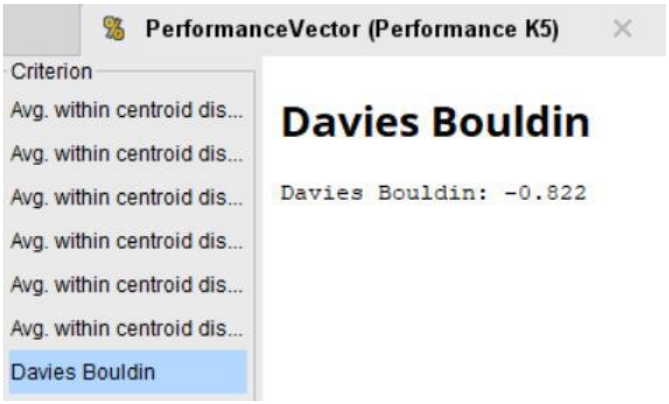
Dalam Indikator ini akan dilihat jarak antar cluster satu dengan lainnya benar-benar berjarak atau tidak. Makin besar jaraknya maka makin bagus, begitupun sebaliknya.

Hasil Evaluasi Davies Bouldin Index

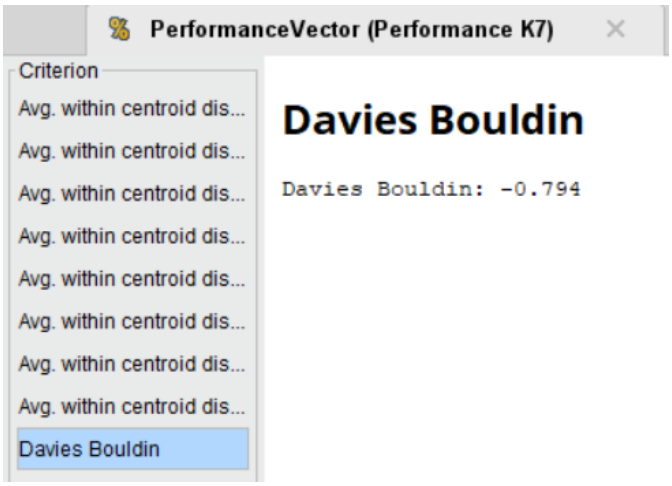
Berikut ini hasil evaluasi dari K-Means 3, K-Means 5 dan K-Means 7:



Gambar 2. Evaluasi Davies Bouldin Index K-Means 3



Gambar 2. Evaluasi Davies Bouldin Index K-Means 5



Gambar 2. Evaluasi Davies Bouldin Index K-Means 7

p-ISSN: 2502-3691
e-ISSN: 2502-4361

Dari hasil di atas ini dapat dilihat bahawa K-Means dengan jumlah cluster 7 mempunyai nilai yang lebih kecil dibanding yang lainnya sehingga dapat disimpulkan bahwa jumlah cluster inilah yang paling baik performance nya sehingga dapat digunakan untuk proses clustering data pelanggannya.

Hasil K-Means Clustering dengan jumlah Cluster 7

1. **Tabel Centroid**

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6
Umur	32.862	25.522	41.647	56.156	44.143	27.316	32.200
Pemasukan (kRp)	78.552	26.304	88.735	53.378	25.143	57.500	109.700
Skor Pengeluaran (1-100)	82.172	78.565	16.765	49.089	19.524	49.447	82

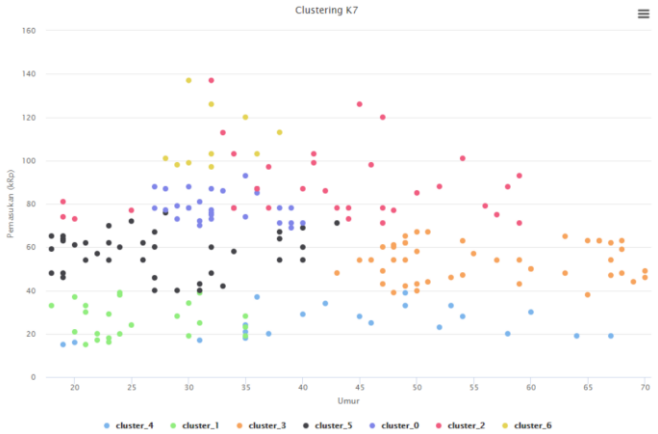
Dari tabel ini sebenarnya sudah bisa dilihat bahwa pengeluaran tertinggi ada di cluster 0 dan cluster 6 sedangkan untuk pemasukan tertinggi ada di cluster 6 dan cluster 2.

2. **Tabel Statistic**

Name	Type	Missing	Statistics		
id	Integer	0	Min: 1	Max: 200	Average: 100.500
cluster	Nominal	0	Least: cluster_6 (10)	Most: cluster_3 (45)	Values: cluster_3 (45), cluster_5 (38), ... [5 more]
Umur	Integer	0	Min: 18	Max: 70	Average: 38.850
Pemasukan (kRp)	Integer	0	Min: 15	Max: 137	Average: 60.560
Skor Pengeluaran (1-100)	Integer	0	Min: 1	Max: 99	Average: 50.200

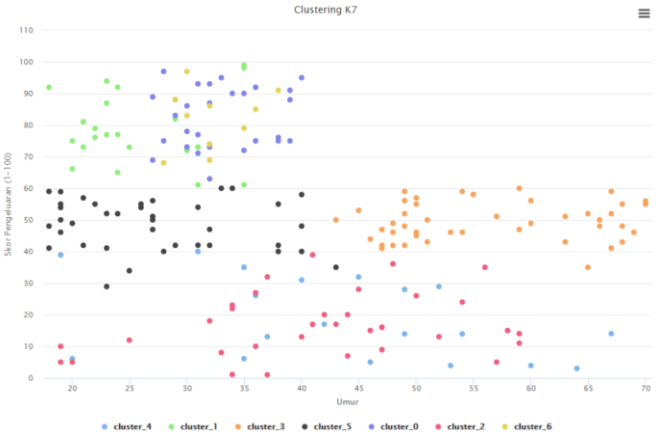
Dari tabel ini dapat kita lihat ringkasan dari data pelanggan baik dari minimal sampai maksimal dan rata-rata datanya. Selain itu juga dapat kita lihat bahwa tidak ada data yang hilang pada field-field-nya.

3. **Hasil Clustering Berdasarkan Umur dan Pemasukan (Income) nya**



Dari hasil ini dapat dilihat bahwa yang mempunyai pemasukkan tinggi ada di cluster 2 dan 6 dengan rentang usia antara 19-59 tahun sedangkan yang mempunyai penghasilan rendah ada di Cluster 1 dan 2 dengan rentang usia dari 19-67 tahun.

4. **Hasil Clustering Berdasarkan Umur dan Pengeluarannya**



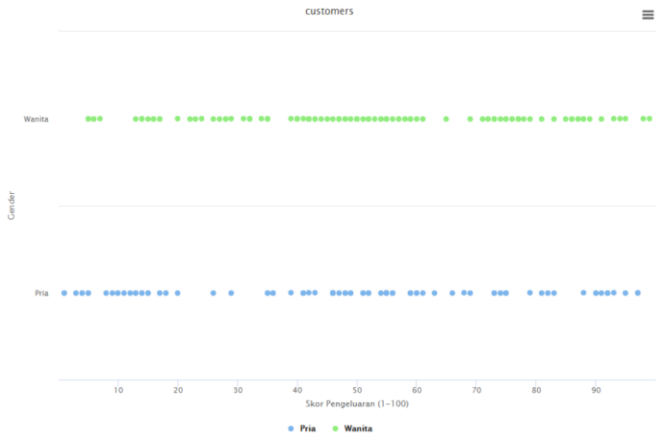
Berdasarkan hasil di atas ini dapat dilihat bahwa cluster dengan pengeluaran terbanyak ada di cluster 0, cluster 1 dan cluster 6 dengan rentang usia antara 18-40 tahun sedangkan cluster dengan pengeluaran sedikit ada di cluster 2 dan cluster 4 dengan rentang usia antara 19-67 tahun.

5. **Hasil Clustering Berdasarkan Gender dan Pemasukannya (Income)**

Dari hasil ini dapat dilihat bahwa yang mempunyai penghasilan tertinggi adalah Pria.

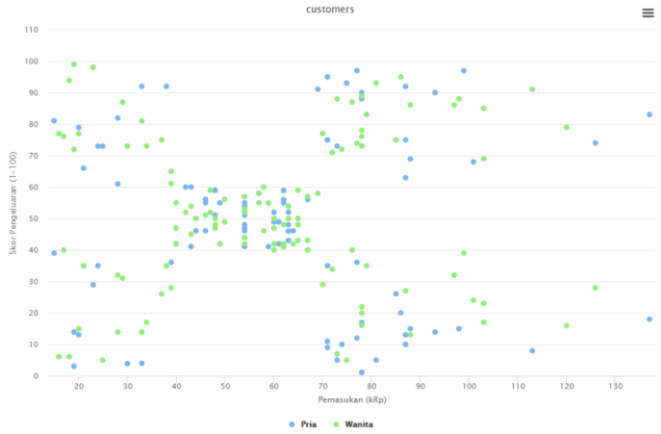
6. **Hasil Clustering Berdasarkan Gender dan Pengeluarannya**

p-ISSN: 2502-3691
e-ISSN: 2502-4361



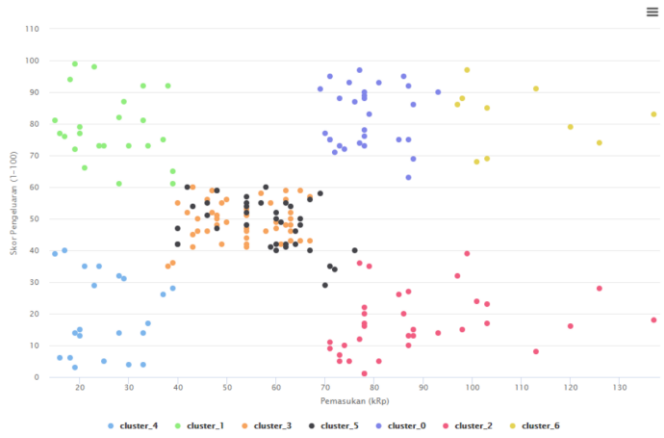
Dari hasil ini dapat dilihat bahwa yang mempunyai penghasilan tertinggi adalah Pria.

7. Hasil Clustering berdasarkan Gender, Pemasukan (Income) dan pengeluarannya



Hasil Clustering ini digunakan untuk memperjelas hasil clustering pada poin (5) dan (6) di atas.

8. Clustering Pelanggan Potensial



Dari hasil clustering ini dapat dilihat bahwa pelanggan potensial ada di Cluster 6 dan Cluster 0. Hal ini ditandai dengan jumlah pemasukan (Income) mereka yang tinggi dan tingkat pengeluarannya juga tinggi.

Untuk Cluster 1 sebenarnya pengeluarannya juga tinggi namun pemasukan rendah. Ini sebenarnya tidak ada masalah bagi penjual karena pada intinya daya beli mereka tinggi namun Cluster ini tidak penulis kategorikan ke pelanggan potensial namun dimasukkan ke pelanggan ceroboh dan berpotensi menimbulkan masalah ke depannya seperti suka utang, kredit, bon dan lain sebagainya untuk memenuhi hasrat membeli namun tidak punya cukup uang.

Di sini yang menarik adalah cluster 3 dan 5 dimana seimbang antara pemasukan dan pengeluarannya standar. Dengan analisa yang lebih dalam lagi di cluster ini punya potensi menjadi pelanggan potensial.

V. KESIMPULAN

Dari hasil penelitian yang diperoleh menggunakan Algoritma K-Means dalam clustering data pelanggan dapat memberikan hasil segmentasi pelanggan yang sesuai dengan harapan, sehingga penjual dapat dengan mudah memahami karakteristik pelanggannya berdasarkan clusternya, pemasukannya, pengeluarannya maupun umur dan gendernya. Kemudian dengan evaluasi Davies Boildin juga memberikan solusi pemilihan jumlah cluster yang tepat sehingga performanya lebih optimal dan menghasilkan hasil segmentasi pelanggan yang lebih akurat.

VI. DAFTAR PUSTAKA

[1] F. Nasari and C. J. M. Sianturi, “Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat,” *CogITo Smart J.*, vol. 2, no. 2, pp. 108–119, 2016, doi: 10.31154/cogito.v2i2.19.108-119.

[2] B. E. Adiana, I. Soesanti, and A. E. Permanasari, “Analisis Segmentasi Pelanggan Menggunakan Kombinasi Rfm Model Dan Teknik Clustering,” *J. Terap. Teknol. Inf.*, vol. 2, no. 1, pp. 23–32, 2018, doi: 10.21460/jutei.2018.21.76.

[3] A. Situmorang, I. Rusilpan, and C. Juliane, “Analisa dan Penerapan Metode Algoritma K-Means Clustering Untuk Mengidentifikasi Rekomendasi Kategori Baru Pada List Movie IMDb,” vol. 6, pp. 2171–2179, 2022, doi: 10.30865/mib.v6i4.4729.

[4] C. D. Rumiarti and I. Budi, “Customer Segmentation for Customer Relationship Management on Retail Company: Case Study PT Gramedia Asri Media,” *J. Sist. Inf.*, vol. 13, no. 1, p. 1, 2017, doi: 10.21609/jsi.v13i1.525.

[5] Maryana, A. Sugianto, Nurmalasari, and A. Ester, “Penerapan Metode Clustering Dengan Algoritma K-Means Untuk Segmentasi Pelanggan Pada CV. Sinar Surya,” *Inti Nusa Mandiri*, vol. 13, no. 1, pp. 39–44, 2018.

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

[7] “STAKEHOLDER PARIWISATA | BINUS UNIVERSITY MALANG | Pilihan Universitas Terbaik di Malang.” [Online]. Available: https://binus.ac.id/malang/2021/08/stakeholder-

p-ISSN: 2502-3691
e-ISSN: 2502-4361

pariwisata/

[8] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono, and C. Lauw, "Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Linear Regression," *J. Nas. Teknol. dan Sist. Inf.*, vol. 7, no. 1, pp. 8–17, 2021, doi: 10.25077/teknosi.v7i1.2021.8-17.

[9] B. Harahap, "Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung)," *Reg. Dev. Ind. Heal. Sci. Technol. Art Life*, pp. 394–403, 2019.

[10] A. A. Az-zahra, A. F. Marsaoly, I. P. Lestyani, R. Salsabila, and W. O. Z. Madjida, "Penerapan Algoritma K-Modes Clustering Dengan Validasi Davies Bouldin Index Pada Pengelompokkan Tingkat Minat Belanja Online Di Provinsi Daerah Istimewa Yogyakarta," *J. MSA (Mat. dan Stat. serta Apl.)*, vol. 9, no. 1, p. 24, 2021, doi: 10.24252/msa.v9i1.18555.