

Analisis Pelanggan Toko Menggunakan Metode Klustering K-Means.pdf

by Raisa Turnitin

Submission date: 06-Jun-2024 10:40AM (UTC+0530)

Submission ID: 2395431186

File name: Analisis_Pelanggan_Toko_Menggunakan_Metode_Klustering_K-Means.pdf (446.54K)

Word count: 1715

Character count: 11453

Analisis Pelanggan Toko Menggunakan Metode Klustering K-Means

¹ st Najwa Afifi Situmorang
Program Studi Ilmu Komputer
Universitas Sumatera Utara
Medan, Indonesia
najwaafifi121@gmail.com

² nd Jovinka Aphellia Salva
Program Studi Ilmu Komputer
Universitas Sumatera Utara
Medan, Indonesia
jovinkav@gmail.com

³ rd Dewi Sartika Br Ginting S.Kom.,
M.Kom
Program Studi Ilmu Komputer
Universitas Sumatera Utara
Medan, Indonesia
dewidintingdg90@gmail.com

Abstract—Dalam artikel ini, algoritma K-Means digunakan untuk menganalisis kelompok pelanggan di sebuah toko untuk mengidentifikasi segmen pelanggan yang berbeda berdasarkan parameter seperti usia, pendapatan tahunan, dan perkiraan pengeluaran. Tujuan dari penelitian ini adalah untuk menciptakan strategi pemasaran yang lebih tepat sasaran dengan menggunakan hasil clustering. Jumlah set yang ideal dapat ditentukan dengan menggunakan teknik elbow dan silhouette. Hasil analisis menunjukkan terdapat empat kelompok utama pelanggan dengan karakteristik berbeda. Karakteristik tersebut kemudian digunakan untuk menciptakan strategi pemasaran yang paling sesuai.

Keywords—Klustering, K-Means, Pemasaran, Analisis Pelanggan, Elbow Method, Silhouette Method.

I. PENDAHULUAN

Di era digital yang berkembang pesat, perusahaan menghadapi tantangan yang semakin kompleks dalam memahami perilaku dan preferensi konsumen. Solusi penting untuk masalah ini adalah segmentasi pelanggan. Hal ini memungkinkan perusahaan untuk menargetkan segmen pelanggan secara spesifik dan efektif. Melalui segmentasi yang efektif, perusahaan tidak hanya dapat mengidentifikasi kebutuhan dan keinginan konsumen, namun juga menciptakan kampanye pemasaran yang lebih tepat sasaran dan efektif.

Pengelompokan adalah salah satu metode segmentasi pelanggan yang paling umum. Algoritma k-means adalah metode populer untuk mengelompokkan data berdasarkan kesamaannya. Data diklasifikasikan ke dalam kelompok-kelompok tertentu dengan karakteristik yang sebanding dalam setiap kelompok. Proses ini dapat mengidentifikasi pola dan tren yang tidak dapat dilihat melalui analisis data tradisional. Sebanding dalam setiap kelompok. Proses ini dapat mengidentifikasi pola dan tren yang tidak dapat dilihat melalui analisis data tradisional.

Data yang paling umum digunakan dalam penelitian ini adalah usia, pendapatan tahunan, dan pendapatan konsumen. Faktor ini dipilih karena pentingnya dalam menentukan perilaku konsumen. Data disiapkan untuk analisis cluster menggunakan teknik pra-pemrosesan yang sesuai.

Menemukan cluster yang sesuai merupakan langkah penting dalam proses K-Means. Dalam penelitian ini, kami menggunakan metode elbow dan silhouette untuk memperkirakan dan menentukan jumlah cluster yang optimal. Metode elbow membantu memutuskan kapan suatu kelompok bergabung dengan kelompok lain. Di sisi lain, fungsi silhouette memperkirakan seberapa mirip satu kelompok dengan kelompok lainnya, yang mencerminkan jumlah informasi yang dibagikan.

Hasil penelitian menunjukkan bahwa terdapat empat kelompok utama dengan karakteristik berbeda. Setiap kelompok budaya memiliki kebutuhan dan preferensi budaya. Bisnis dapat menggunakan fitur-fitur hebat ini untuk mengembangkan strategi pemasaran yang lebih bertarget dan efektif. Misalnya, pengecer yang menjual produk mahal dalam jumlah besar mungkin yakin akan kualitasnya, sementara merek berbiaya rendah dan berkinerja tinggi cenderung menurunkan harga.

Secara keseluruhan, penelitian ini menunjukkan bahwa analisis cluster menggunakan K-Means tidak hanya membantu pelanggan yang lebih baik berdasarkan segmen, tetapi juga membantu perusahaan mengembangkan strategi pemasaran yang lebih baik untuk memenuhi kebutuhan setiap segmen pelanggan.

II. METODOLOGI

A. Pengumpulan dan Pemahaman Data

Ciri-ciri utama adalah umur, tingkat pendapatan dan pengeluaran tahunan. Data ini mewakili demografi dan kebiasaan belanja pelanggan inti toko. Gunakan grafik untuk menjelajahi data mendasar dan lebih memahami distribusi data dan hubungan komponen.

B. Preprocessing Data

Pemrosesan awal data mencakup beberapa langkah utama:

Pembersihan data berarti menangani nilai yang hilang dan menghapus duplikat.

Petunjuk: Memastikan setiap bagian sama pentingnya. Oleh karena itu, digunakan metode deviasi standar minimum-maksimum untuk menentukan data. Hal ini sangat penting terutama untuk metode K-means, yang sensitif terhadap parameter statistik.

Konversi Data: Temukan data rahasia dalam urutan numerik bila memungkinkan.

C. Pemilihan Jumlah Kluster

Salah satu langkah penting dalam clustering K-means adalah menentukan jumlah cluster yang optimal. Prosesnya adalah sebagai berikut;

Metode elbow: Nilai inertia, yang merupakan jumlah kuadrat jarak ke setiap titik, ditampilkan untuk jumlah kelompok yang berbeda. Titik di mana redundansi mulai berkurang adalah ketika jumlah cluster yang optimal telah dipilih.

Penilaian silhouette: Penilaian silhouette digunakan untuk meningkatkan efisiensi pengomposisian. Jika eksponennya

mendekati 1, maka data tersebut diklasifikasikan dengan baik.

D. Implementasi K-Means

Implementasi Langkah demi Langkah Algoritma K-Means:

Inisialisasi Centroid: Anda harus memiliki titik fokus K untuk memulai titik fokus.

Penugasan Cluster: Tetapkan titik data ke cluster berdasarkan jarak Euclidean.

Hitung Ulang Titik Centroid: Memperbarui posisi titik tengah menggunakan rata-rata seluruh titik dalam grup.

Iterasi: Menghitung titik tengah dan menetapkan cluster diulangi hingga konvergensi atau setidaknya titik tengah serupa tercapai.

E. Evaluasi Model

Hasil evaluasi menggunakan model K-means:

Visualisasi Cluster: Pisahkan dan isolasi kelompok menggunakan ruang 2D atau 3D.

Analisis profil Cluster: Untuk memahami profil setiap segmen pelanggan, analisis setiap kelompok berdasarkan karakteristik utama seperti usia rata-rata, pendapatan, dan pengeluaran.

III. PEMBAHASAN

Dataset ini berisi 16.000 baris data, atau 2000 baris x 8 kolom. Pelanggan direpresentasikan sebagai string yang berisi berbagai informasi demografi dan perilaku. Kumpulan data ini digunakan untuk analisis cluster, yang membantu menentukan segmen pelanggan berdasarkan karakteristik dan perilakunya. Ini membantu Anda menciptakan strategi pemasaran yang efektif dan layanan pelanggan yang dipersonalisasi.

RangeIndex: 2000 entries, 0 to 1999

Data columns (total 8 columns):

| # | Column | Non-Null Count | Dtype |
|---|------------------------|----------------|--------|
| 0 | CustomerID | 2000 non-null | int64 |
| 1 | Gender | 2000 non-null | object |
| 2 | Age | 2000 non-null | int64 |
| 3 | Annual Income (\$) | 2000 non-null | int64 |
| 4 | Spending Score (1-100) | 2000 non-null | int64 |
| 5 | Profession | 1965 non-null | object |
| 6 | Work Experience | 2000 non-null | int64 |
| 7 | Family Size | 2000 non-null | int64 |

dtypes: int64(6), object(2)

A. Deskripsi Data

- Jumlah nomor identifikasi unik untuk setiap pelanggan disimpan dalam CustomerID (2000 non-null, int64).
- Jenis kelamin pelanggan: Jenis kelamin pelanggan (2000 non-null, objek), seperti "Laki-laki" atau "Wanita".
- Usia pelanggan: Usia pelanggan pada tahun 2000 (non-null, int64).

- Tahunan Pendapatan (\$): Pendapatan pelanggan tahunan dalam dolar (2000 non-null, int64).
- Spending Score (1-100): Skor pengeluaran berdasarkan perilaku belanja pelanggan, dengan nilai mulai dari 1 hingga 100 (2000 non-null, int64).
- Profesi: Pekerjaan pelanggan, dengan beberapa data yang hilang (1965 non-null, objek).
- Pengalaman Kerja: Pengalaman kerja pelanggan pada tahun 2000 (non-null, int64)
- Ukuran keluarga: Ukuran keluarga klien (2000 non-null, int64).

| | CustomerID | Gender | Age | Annual Income (\$) | Spending Score (1-100) | Profession | Work Experience | Family Size |
|---|------------|--------|-----|--------------------|------------------------|---------------|-----------------|-------------|
| 0 | 1 | Male | 19 | 15000 | 39 | Healthcare | 1 | 4 |
| 1 | 2 | Male | 21 | 35000 | 81 | Engineer | 3 | 3 |
| 2 | 3 | Female | 20 | 86000 | 6 | Engineer | 1 | 1 |
| 3 | 4 | Female | 23 | 59000 | 77 | Lawyer | 0 | 2 |
| 4 | 5 | Female | 31 | 38000 | 40 | Entertainment | 2 | 6 |

B. K-means

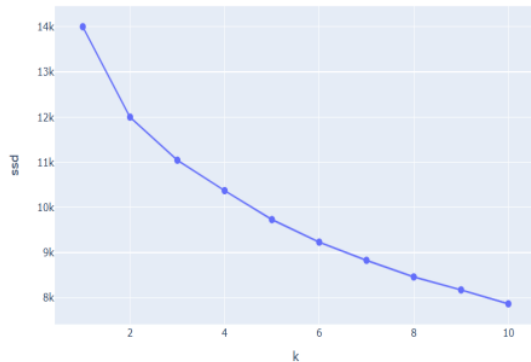
"k-means clustering" digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang disebut "cluster" untuk memastikan bahwa data dalam satu cluster memiliki kemiripan yang tinggi dan perbedaan yang signifikan dengan data di cluster lainnya. Algoritma ini bekerja dengan menghitung centroid atau centroid cluster dan kemudian mengatur setiap titik data ke dalam cluster.

1. Grafik Elbow

Metode Elbow menggunakan algoritma K-means clustering. Cara ini memerlukan langkah yang berbeda-beda dan bergantung pada nomor komponen, nomor komponen yang digunakan, dan sudut layar. Persegi standar atau cluster sico-square (VCSS) adalah perbedaan statistik antara data dan data.

Metode Elbow digunakan untuk memaksimalkan cluster dan meminimalkan VCSS di setiap segmen. Namun, kini saatnya memikirkan tentang VCSS yang merupakan agen penyebab lebah. Jumlah unit yang digunakan ditunjukkan dengan "siku" atau jumlah unit yang digunakan.

Elbow Method



2. Silhouette Method

Metode Silhouette adalah cara untuk mengevaluasi kinerja algoritma sekuensing. Ini digunakan untuk menghitung tingkat kesamaan kumpulan data dalam clusternya sendiri dan antar cluster lainnya. Setiap rekor diberi skor dari -1 hingga 1:

Jika suatu kelompok data mendapat skor 1, maka skor tersebut berada pada kelompoknya sendiri. Namun jika 0, maka tepinya berada di antara dua grup.

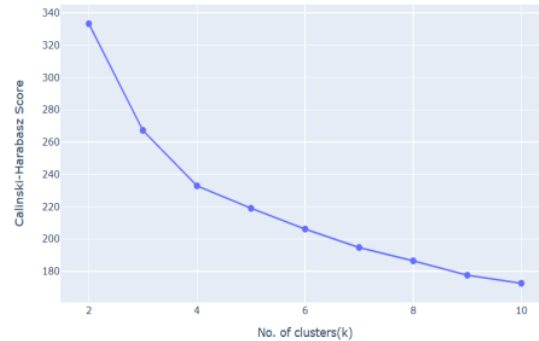
Entri data bisa berada di kolom kesalahan jika diberi nilai -1. Skor untuk solusi cluster ditentukan dengan menjumlahkan poin skor untuk setiap record dalam kumpulan data. Tingkatkan solusi dokumentasi Anda dengan IPK Silhouette.

Untuk setiap segmen, metode pencocokan menghitung skor, k. Jumlah segmen yang optimal dipilih berdasarkan skor tinggi dari siluet. Metode siluet dapat digunakan bersamaan dengan metode manual untuk menentukan jumlah segmen yang optimal untuk pasar tertentu.

Sebagai referensi, skor dapat dihitung sebagai:

Untuk mengukur jarak antara i dan semua titik lain dalam cluster, hitunglah nilai (i). Ukur juga jarak antara titik tersebut dan semua titik lain dalam kumpulan batas. Nilai yang ditampilkan adalah b(i). Bandingkan skor gambar dengan $s(i) = (b(i) - a(i)) / \max(a(i), b(i))$. Untuk hasil korelasi, skor totalnya merupakan rata-rata seluruh skor pada setiap dataset.

Calinski-Harabasz Index



3. Calinski-Harabasz Index

Indeks kepadatan relatif, atau indeks Calinski-Harabasz, adalah ukuran penyaringan statistik yang menghitung rasio kepadatan di suatu wilayah tertentu terhadap wilayah yang tersebar.

Jarak antara titik-titik individu dalam kelompok dan pusat disebut simpangan baku dalam kelompok. Jarak antara pusat tiap kelompok disebut simpangan baku.

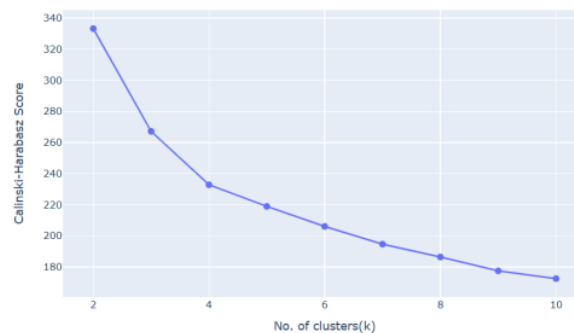
Rumusnya:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

dimana n adalah jumlah seluruh data, B (k) adalah selisih antar kelompok, W (k) adalah selisih antar kelompok, dan CH (k) adalah indeks Kalinsky-Harabasz untuk nilai tertentu k (jumlah kelompok)).

Rasio penyebaran dalam kelompok terhadap dalam kelompok diperkirakan akan meningkatkan indeks Calinski-Harabasz. Nilai indeks Calinski-Harabasz ditambahkan untuk memperoleh cluster yang lebih besar

Calinski-Harabasz Index



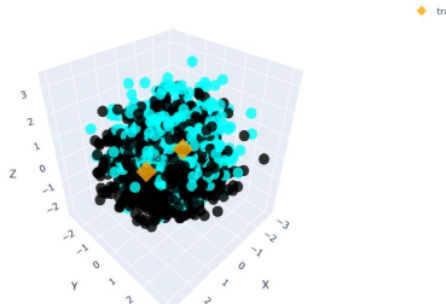
C. Visualisasi

1. VISUALISASI 3D:

Dengan bantuan presentasi tiga dimensi, kami menggambarkan distribusi data tiga dimensi. Hal ini menciptakan tiga perbedaan yang berbeda. Survei ini menyajikan angka pendapatan dan biaya tahunan serta pengalaman pelanggan dalam pencitraan 3D. Gambar tiga dimensi memperlihatkan bentuk dan kelompok yang tidak terlihat dalam tampilan dua dimensi. Memfasilitasi definisi distribusi informasi tiga dimensi dan konstruksi cluster k-means.

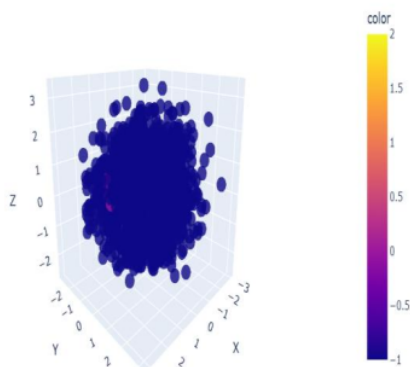
- Visualisasi K-means

K Means Clustering Visualization



- Klustering DBSCAN

DBSCAN Clustering(3 Clusters)

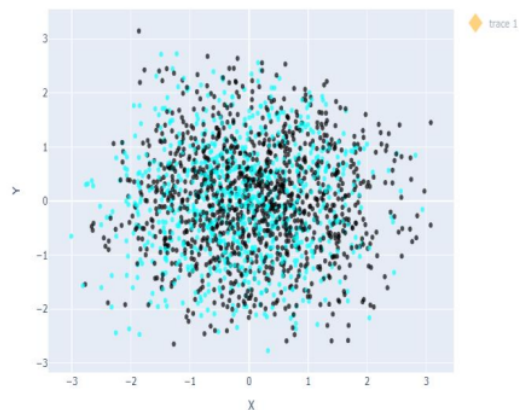


2. VISUALISASI 2D:

Penelitian ini menggunakan visual 2D berukuran besar untuk mengkaji sebaran data. Beberapa variabel dibandingkan untuk membandingkan hubungan antar variabel dan untuk mengidentifikasi kelompok mana yang didukung. Misalnya, pengakuan pendapatan berkaitan dengan jumlah produk yang terjual. Denah 2D sederhana dan mudah dicocokkan dengan peta 3D, namun detailnya terlihat jelas.

- Visualisasi K-means

K Means Clustering Visualization



- Klustering DBSCAN

DBSCAN Clustering



IV. KESIMPULAN

K-means jurnal ini digunakan untuk memperkirakan pelanggan berdasarkan usia, pendapatan tahunan, dan perkiraan pengeluaran. Prosesnya diawali dengan pengumpulan data dan preprocessing. Proses ini melibatkan pembersihan, penyimpanan, dan konversi data non-numerik menjadi bentuk numerik. Para peneliti menggunakan metode siku dan siluet untuk mengidentifikasi jumlah cluster yang optimal. Tren dan tren data pelanggan ditentukan oleh algoritma K-Means.

Temuan menunjukkan bahwa masing-masing dari empat kelompok konsumen primer memiliki kualitas yang berbeda. Pola konsumsi dan demografi tiap cluster berbeda. Klien yang lebih muda dengan pendapatan dan pengeluaran yang moderat, misalnya, tidak sama dengan klien yang lebih tua dengan pendapatan dan pengeluaran yang sama. Temuan ini membantu dunia usaha untuk lebih memahami

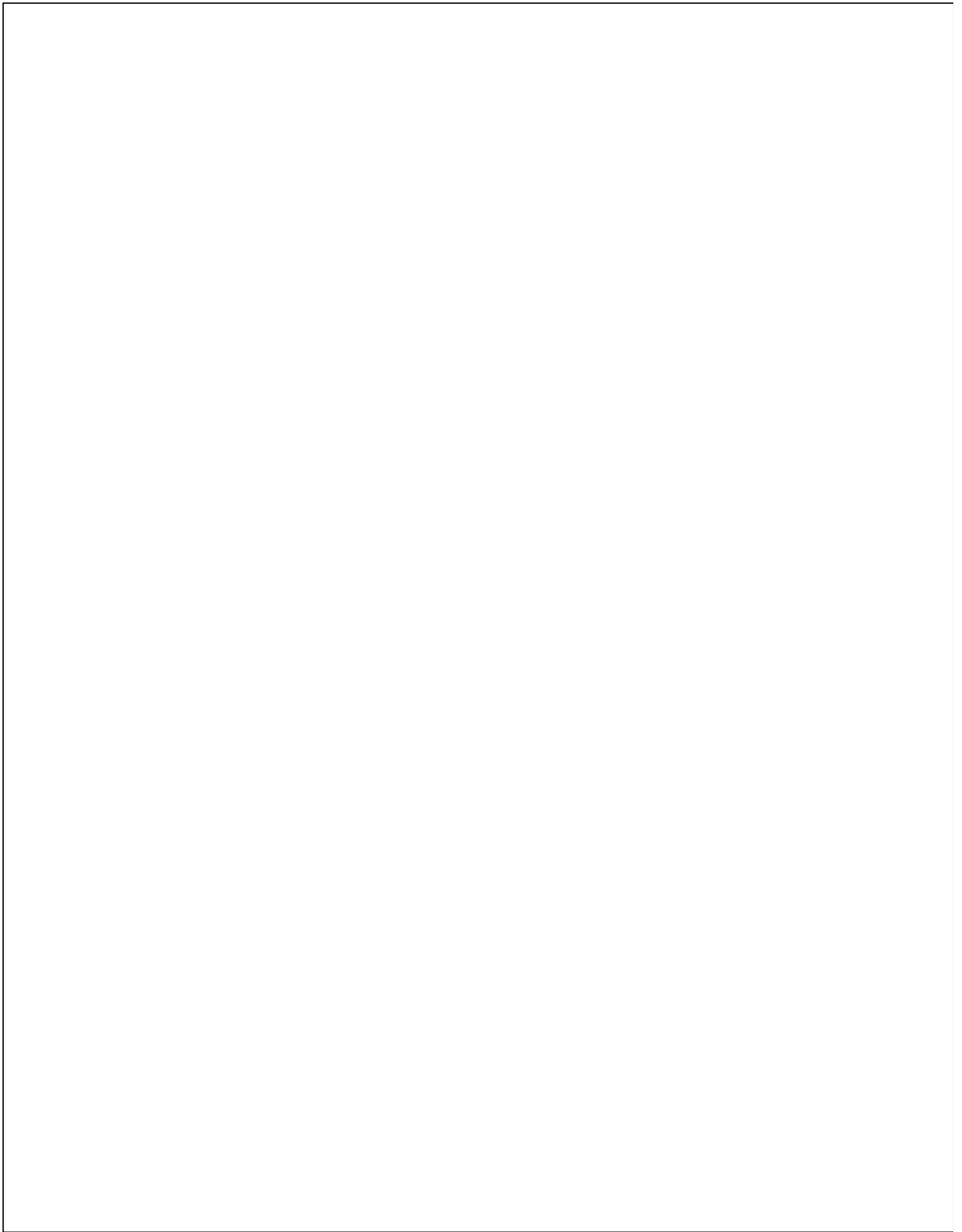
kebutuhan dan preferensi setiap segmen konsumen, sehingga memungkinkan mereka mengembangkan rencana pemasaran yang lebih fokus dan sukses.

Secara keseluruhan, penelitian ini menunjukkan bahwa metode K-Means untuk segmentasi pelanggan dapat memberikan informasi yang berguna tentang preferensi dan perilaku pelanggan. Dengan menggunakan informasi ini, perusahaan dapat menciptakan strategi pemasaran yang lebih sesuai untuk setiap segmen pelanggan, sehingga meningkatkan kepuasan pelanggan dan, pada akhirnya, pendapatan. Metode ini menunjukkan pentingnya analisis data dalam menciptakan rencana bisnis yang lebih cerdas dan efektif.

REFERENCES

- [1] Hua H Y, Zhao H C. Application of Clustering Algorithms in Bank Customer Segmentation [J].
- [2] Syakur, M. A., et al. Integration k-means clustering method and elbow method for identification of the
- [3] Shop Customer Clustering. <https://www.kaggle.com/code/utkarshsaxenadn/shop-customer-clustering/notebook#Data-Visualization>

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.



Analisis Pelanggan Toko Menggunakan Metode Klustering K-Means.pdf

ORIGINALITY REPORT

6%

SIMILARITY INDEX

4%

INTERNET SOURCES

3%

PUBLICATIONS

1%

STUDENT PAPERS

PRIMARY SOURCES

| | | |
|---|--|-----|
| 1 | publikasi.dinus.ac.id Internet Source | 1% |
| 2 | www.androtechno.my.id Internet Source | 1% |
| 3 | fasilkom-ti.usu.ac.id Internet Source | 1% |
| 4 | Submitted to Sriwijaya University Student Paper | 1% |
| 5 | Gilang Fajriansyah. "ANALISIS DAFTAR PEMILIH TETAP PADA HASIL REKAPITULASI KPU BERDASARKAN USIA MENGGUNAKAN ALGORITMA K-MEANS (STUDI KASUS : KOTA BANDAR LAMPUNG)", Electrician, 2021 Publication | 1% |
| 6 | www.scribd.com Internet Source | 1% |
| 7 | Anindya Khrisna Wardhani. "K-MEANS ALGORITHM IMPLEMENTATION FOR CLUSTERING OF PATIENTS DISEASE IN KAJEN | <1% |

CLINIC OF PEKALONGAN", Jurnal Transformatika, 2016

Publication



paketcctvindonesia.blogspot.com
Internet Source

<1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography On