Potensi Big Data menggunakan Algoritma Clustering untuk Segmentasi E-Commerce pada Pelanggan Perusahaan

Angga Pramana Putra Wibowo[1 *], Puti Windrahmatullah[2], Dhea Amelia Putri[3], Renta Siahaan[4], Elilya Octaviani[5] , Smertniki Javid Ahmedthian[6]

[1]Prodi Sains Data, Fakultas Sains, Institut Teknologi Sumatera, Kab. Lampung Selatan, Indonesia
**E-mail: angga.120450084@student.itera.ac.id , puti.120450070@student.itera.ac.id , dhea.122450004@student.itera.ac.id , renta.122450070@student.itera.ac.id , elilya.122450009@student.itera.ac.id , smertniki.122450115@student.itera.ac.id*

Abstrak

Dalam era digital dan persaingan ketat *E-commerce*, pembaruan teknologi diperlukan untuk memahami pelanggan. Segmentasi membantu mengidentifikasi kelompok pelanggan serupa. Penggunaan teknologi baru seperti machine learning dan data mining penting. Analisis pelanggan melibatkan teknik seperti clustering dan analisis perilaku. Pengambilan keputusan cerdas, seperti sistem rekomendasi, digunakan untuk penawaran promo atau diskon yang sesuai. Evaluasi dan penyesuaian dilakukan untuk memantau kinerja dan respons pelanggan Program loyalitas yang memberikan penghargaan kepada pelanggan untuk pembelian mereka. Contohnya program loyalitas yang memberikan poin untuk setiap pembelian, yang dapat ditukarkan dengan hadiah seperti produk gratis, diskon, atau akses ke acara eksklusif. Penawaran khusus dan diskon untuk pelanggan setia untuk mendorong pembelian lebih lanjut. Seperti pengiriman gratis atau diskon tambahan untuk pembelian berulang. Lakukan survei kepuasan pelanggan untuk memahami faktor yang dapat meningkatkan tingkat kepuasan dan mengurangi tingkat pembatalan.menggunakan algoritma clustering pada big data pelanggan e-commerce tidak hanya membantu perusahaan dalam menyediakan layanan yang lebih baik kepada pelanggan, tetapi juga mengoptimalkan penggunaan sumber daya mereka. Dalam konteks ekonomi mikro, ini semua terkait dengan konsep efisiensi, perilaku konsumen, serta permintaan dan penawaran di pasar.

Kata kunci: Kmeans; Pelanggan*; E-commerce;* Perusahaan; clustering; Algoritma

Abstract

In the digital era and intense competition of E-commerce, technological updates are needed to understand customers. Segmentation helps identify groups of similar customers. The use of new technologies such as machine learning and data mining is important. Customer analysis involves techniques such as clustering and behavioral analysis. Intelligent decision making, such as recommendation systems, are used to offer appropriate promotions or discounts. Evaluations and adjustments are made to monitor customer performance and response Loyalty programs that reward customers for their purchases. For example, loyalty programs award points for every purchase, which can be exchanged for prizes such as free products, discounts, or access to exclusive events. Special offers and discounts for loyal customers to encourage further purchases. Such as free shipping or additional discounts for repeat purchases. Conduct customer satisfaction surveys to understand factors that can increase satisfaction levels and reduce cancellation rates. Using clustering algorithms on e-commerce customer big data not only helps companies provide better services to customers, but also optimizes the use of their resources. In a microeconomic context, this is all related to the concept of efficiency, consumer behavior, and demand and supply in the market.

Keywords: Kmeans; Customer; Ecommerce; Company; clustering; Algorithm

PENDAHULUAN

Dalam era digital yang kompetitif, pembaruan teknologi seperti machine learning dan data mining menjadi penting untuk memahami pelanggan. Segmentasi melalui algoritma clustering memungkinkan perusahaan e-commerce untuk mengidentifikasi kelompok pelanggan serupa dan mengoptimalkan strategi pemasaran. Evaluasi teknologi baru dalam clustering penting untuk memastikan akurasi segmentasi.

Algoritma clustering memainkan peran kunci dalam segmentasi pelanggan, membantu perusahaan dalam:

1. Efisiensi Alokasi Sumber Daya: Mengalokasikan sumber daya dengan lebih baik berdasarkan pola pembelian dan preferensi.
2. Diferensiasi Produk dan Harga: Memahami preferensi pelanggan untuk menyesuaikan produk dan harga.
3. Teori Permintaan dan Penawaran: Menyesuaikan strategi penawaran berdasarkan respons segmen pelanggan.
4. Efisiensi Operasional: Meningkatkan logistik dan manajemen inventaris berdasarkan segmentasi pelanggan.

Dengan demikian, penggunaan big data dan algoritma clustering tidak hanya meningkatkan layanan pelanggan, tetapi juga efisiensi operasional dan strategi bisnis yang lebih efektif.

METODE

Pada metode penelitian terdiri atas beberapa tahapan proses, yakni *prepro-cessing* data yang akan digunakan untuk implementasi lalu menganalisis parameter yang digunakan untuk memproses data dengan melakukan pembagian data yang telah diolah. Pembagian data untuk data *training* dan data *testing* dengan perban- dingan sebesar 80:20. Kemudian merancang sebuah model algoritma.

Dataset

Dataset [Online Retail II] (https://archive.ics.uci.edu/dataset/502/online+retail+ii) berisi semua transaksi yang terjadi untuk perusahaan ritel online yang berbasis di Inggris dan terdaftar, tanpa toko fisik, antara tanggal 01/12/2009 dan 09/12/2011. Perusahaan ini utamanya menjual barang-barang hadiah serba guna yang unik. Banyak pelanggan perusahaan ini adalah pedagang grosir.

Dataset terdiri dari 525461 baris 8 fitur yaitu:

- Invoice: Nomor Faktur. Nominal. Nomor integral 6 digit yang diberikan secara unik untuk setiap transaksi. Jika kode ini dimulai dengan huruf 'C', itu menandakan pembatalan.
- StockCode: Kode Produk (barang). Nominal. Nomor integral 5 digit yang diberikan secara unik untuk setiap produk yang berbeda.
- Description: Nama Produk (barang). Nominal.
- Quantity: Jumlah dari setiap produk (barang) per transaksi. Numerik.
- InvoiceDate: Tanggal dan waktu Faktur. Numerik. Hari dan waktu ketika transaksi dihasilkan.
- UnitPrice: Harga per unit. Numerik. Harga produk per unit dalam mata uang pound sterling (£).
- Customer ID: Nomor Pelanggan. Nominal. Nomor integral 5 digit yang diberikan secara unik untuk setiap pelanggan.
- Country: Nama Negara. Nominal. Nama negara tempat pelanggan tinggal.

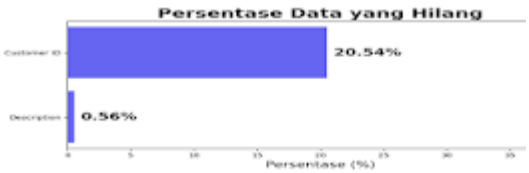
2.2 Data Cleaning

Bagian Data Cleaning fokus pada membersihkan dataset dari masalah kualitas data. Beberapa langkah yang dilakukan di sini termasuk:

- Menghitung dan memvisualisasikan persentase data yang hilang dalam setiap kolom.

- Menghapus baris yang memiliki data yang hilang pada kolom 'Customer ID' dan 'Description' karena keduanya penting dan tidak dapat digantikan.
- Menghapus data duplikat.
- Mengidentifikasi dan menghapus anomali pada kolom 'StockCode' yang memiliki karakter yang tidak sesuai.
- Membersihkan deskripsi produk dengan menghapus tanda baca dan mengubahnya menjadi huruf besar.
- Menghapus baris dengan harga 0.

2.2.1 Missing Value



Gambar 1

Missing data pada Customer ID dan Description akan dihapus Missing data pada CustomerID dihapus karena CustomerID merupakan fitur yang penting dan tidak bisa digantikan. Missing data pada description dihapus karena tidak terlalu banyak missing value dan tidak bisa dilakukan penyesuaian dengan StockCode karena terdapat ketidak konsistenan berdasarkan deskripsi statistik sebelumnya, yaitu dengan persentase data yang hilang customer 20.54% dan product 0.56%

2.2.2 Data Duplikat

Digunakan untuk memproses dataset faktur. Pertama, dilakukan pengecekan terhadap jumlah baris duplikat dalam dataset, dan ditemukan bahwa ada 6771 baris yang merupakan duplikat. Selanjutnya, baris-baris duplikat ini dihapus dari dataset, dan setelah proses penghapusan, dimensi dataset saat ini adalah 410763 baris dan 8 kolom. Penghapusan baris duplikat dilakukan untuk menghindari bias dan meningkatkan konsistensi serta kualitas data.

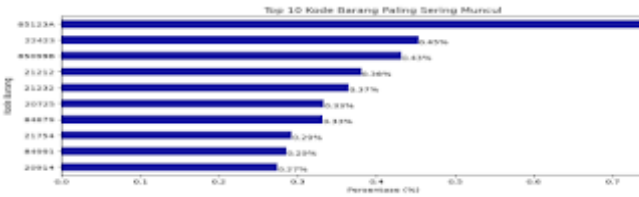
2.2.3 Transaksi Yang dibatalkan

Dalam analisis ini, telah dibuat kolom baru yang menyatakan status transaksi, di mana transaksi dengan kode Invoice yang diawali huruf 'C' dianggap sebagai transaksi yang dibatalkan. Setelah itu, dilakukan analisis statistik deskriptif terhadap transaksi yang dibatalkan, mengungkapkan bahwa harga barang yang dibatalkan cukup bervariasi tanpa kecenderungan tertentu terhadap suatu barang.

Selanjutnya, dilakukan perhitungan persentase pembatalan transaksi terhadap keseluruhan dataset. Hasilnya menunjukkan bahwa persentase transaksi yang dibatalkan dalam dataset tersebut adalah sebesar 2.39%, memberikan gambaran bahwa sebagian kecil dari total transaksi mengalami pembatalan.

2.2.4 Data Anomali

Kode Barang Anomali
Post
D
M
C2
BANK CHARGES
TEST001
TEST002
PADS
ADJUST
ADJUST2
SP1002



Gambar 2

Terdapat anomali pada StockCode dengan total 11 kode menyimpang dari aturan penulisan StockCode yang seharusnya terdiri dari 5 karakter. Persentase kode barang yang anomali dalam dataset yaitu 0.44%. Kode barang anomaly akan dihapus terdapat ketidakkonsistenan karena tanda baca seperti koma. Hapus tanda baca dan buat semua deskripsi menjadi huruf besar.

2.2.5 Outliers

Gambar 3

Data outliers adalah data yang memiliki nilai yang jauh berbeda dari sebagian besar data lainnya. Data outliers dapat mengganggu hasil analisis segmentasi pelanggan, karena dapat mempengaruhi nilai rata-rata, standar deviasi, dan distribusi data. Oleh karena itu, data outliers akan dihapus dari data yang dipakai untuk segmentasi, dan akan dianalisis secara terpisah dari data inliers (data yang bukan outliers).

Dari gambar yang diberikan, dapat dilihat bahwa garis orange menunjukkan persentase inliers, yaitu 90% dari total data. Hal ini menunjukkan bahwa 90% dari populasi mengikuti pola umum yang sama. Garis putih/orange pendek menunjukkan persebtase outliers, yaitu 10% dari total data. Hal ini menunjukkan bahwa 10% dari populasi tidak mengikuti pola umum yang sama. Penyebab terjadinya outliers dapat berupa kesalahan data, kesalahan pengambilan data, atau faktor lain yang tidak terduga.

2.3 Feature Engineering

Feature Engineering adalah bagian yang berfokus pada pembuatan fitur-fitur baru yang dapat memberikan wawasan lebih dalam tentang data. Beberapa langkah yang dilakukan di sini termasuk:

- Menghitung jumlah hari sejak pelanggan melakukan pembelian terakhir.
- Menghitung total transaksi, total produk yang dibeli, total belanja, dan rata-rata belanja per pelanggan.
- Menghitung total jenis produk yang dibeli oleh setiap pelanggan.
- Menghitung rata-rata jumlah hari antara pembelian berturut-turut.
- Menentukan hari dan jam favorit pelanggan untuk berbelanja.
- Menggambarkan apakah pelanggan berasal dari Inggris atau tidak dengan variabel biner 'Is_UK'.
- Menghitung total pembatalan dan tingkat pembatalan transaksi.
- Menghitung rata-rata belanja per bulan dan deviasi standar pengeluaran bulanan pelanggan.

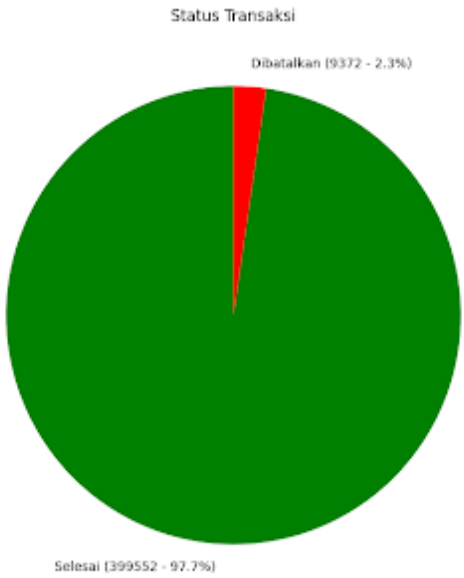
Terdapat 15 Fitur di dapatkan dari fitur Engineering, yaitu :

- Customer ID: Identifier unik untuk membedakan setiap pelanggan.
- Hari_Sejak_Pembelian_Terakhir: Jumlah hari sejak pembelian terakhir pelanggan.
- Total_Transaksi: Total jumlah transaksi yang dilakukan pelanggan.
- Total_Produk_Dibeli: Jumlah total produk yang dibeli pelanggan.
- Total_Belanja: Total uang yang dihabiskan pelanggan.
- Rata_rata_Belanja: Nilai rata-rata setiap transaksi pelanggan.
- Total_Jenis_Produk_Dibeli: Jumlah produk berbeda yang telah dibeli pelanggan.

- Rata_Hari_Antara_Pembelian: Rata-rata hari antara pembelian berturut-turut.
- Hari: Hari pelanggan cenderung berbelanja.
- Jam: Jam pelanggan cenderung berbelanja.
- Is_UK: Variabel biner menunjukkan apakah pelanggan berbasis di Inggris.
- Frekuensi_Pembatalan: Jumlah total transaksi yang dibatalkan oleh pelanggan.
- Tingkat_Pembatalan: Proporsi transaksi yang dibatalkan dari total transaksi.
- Rata_Bulan: Rata-rata pengeluaran bulanan pelanggan.
- Std_Bulan: Deviasi standar pengeluaran bulanan, menunjukkan variabilitas dalam pola pengeluaran pelanggan.

2.4 Eksplorasi Data Analisis (EDA)

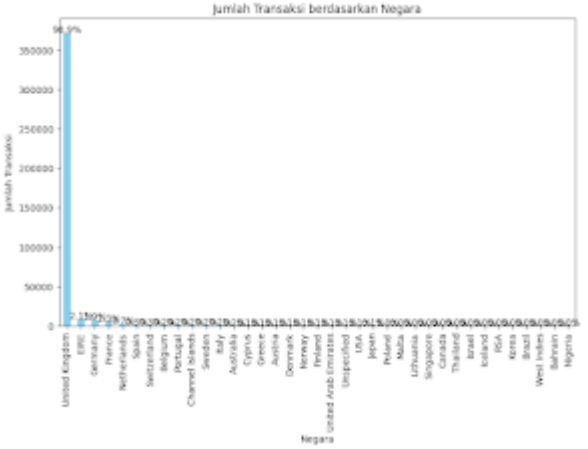
Eksplorasi data atau EDA (*Exploratory Data Analysis)* adalah suatu proses awal yang dilakukan untuk mengidentifikasi pola, proses penemuan anomali, proses uji hipotesis dan pemeriksaan asumsi. Proses EDA dimulai dengan analisis status transaksi



Gambar 4.

Gambar diatas menunjukkan bahwa sebagian besar transaksi (97.7%) berakhir dengan status selesai, sedangkan 2.3% transaksi dibatalkan. Visualisasi menggunakan pie chart memberikan gambaran yang jelas tentang proporsi status transaksi.

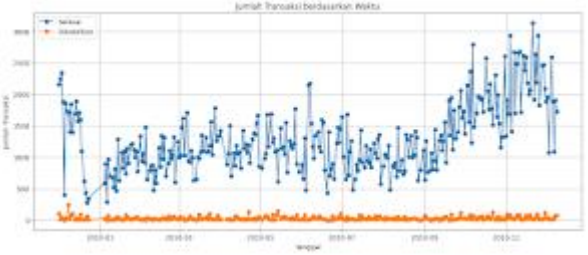
Selanjutnya, dilakukan analisis terhadap negara asal pelanggan.



Gambar 5.

Hasilnya menunjukkan bahwa 90.9% transaksi berasal dari United Kingdom, diikuti oleh Irlandia dengan 2.1% transaksi. Grafik batang digunakan untuk mengilustrasikan jumlah transaksi berdasarkan negara.

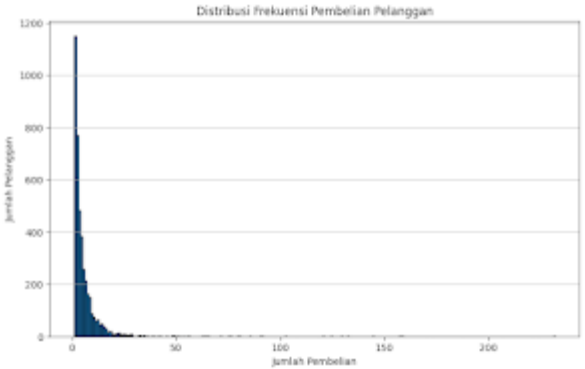
Tren waktu transaksi dieksplorasi dengan membuat line chart.



Gambar 6.

Grafik diatas menunjukkan kenaikan jumlah transaksi dari bulan September 2010 hingga Desember 2010, menunjukkan adanya tren pertumbuhan transaksi.

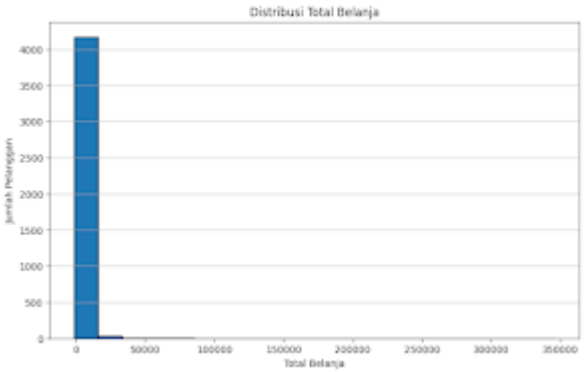
Distribusi frekuensi pembelian pelanggan kemudian dianalisis.



Gambar 7.

Dengan nilai Skewness sebesar 9.5496 yang menandakan skewness positif yang artinya mayoritas pelanggan melakukan pembelian dalam jumlah yang relatif kecil, namun ada beberapa pelanggan (outlier) yang melakukan jumlah pembelian jauh lebih tinggi.

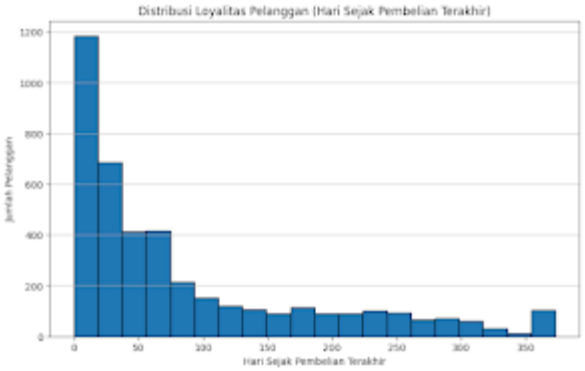
Distribusi Total Belanja



Gambar 8.

Dari grafik diatas dapat disimpulkan Total belanja mayoritas pelanggan relatif kecil. Tetapi terdapat beberapa pelanggan yang telah berbelanja dengan nilai yang cukup besar

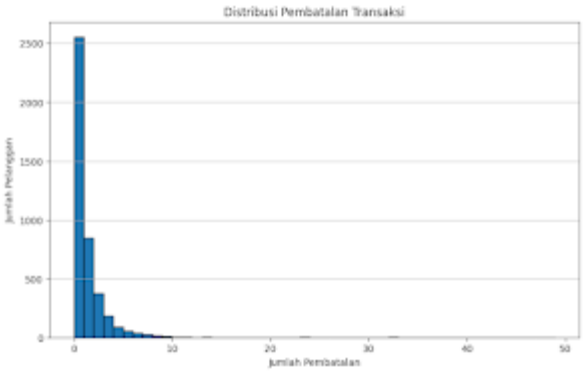
Loyalitas Pelanggan



Gambar 9.

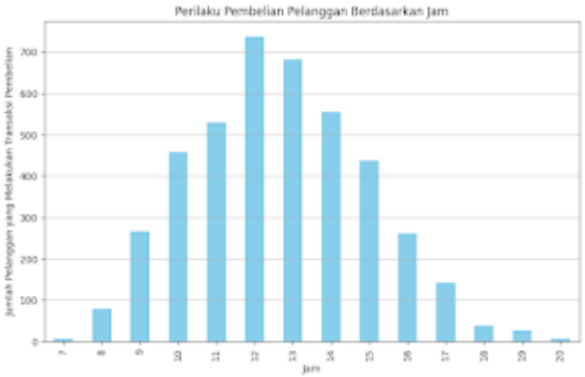
Dilihat dari grafik diatas Mayoritas pelanggan melakukan pembelian secara rutin.

Distribusi Pembatalan Transaksi



Gambar 10.

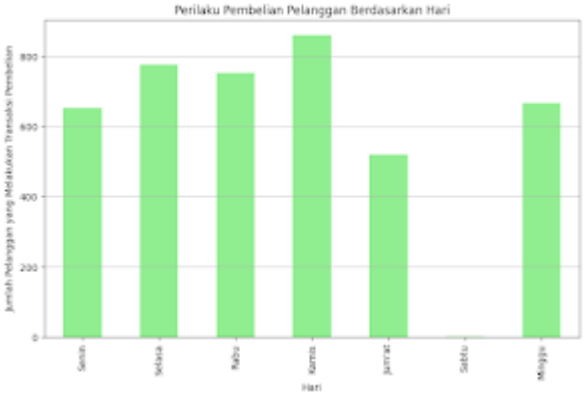
Grafik diatas menunjukkan Sebagian besar pelanggan memiliki jumlah pembatalan yang lebih rendah, namun ada beberapa pelanggan dengan jumlah pembatalan yang sangat tinggi, sehingga menyebabkan perluasan distribusi.



Gambar 11.

Grafik diatas menunjukkan Mayoritas pelanggan lebih suka melakukan pembelian di pagi hingga sore hari, mulai dari jam 9 pagi sampai jam 4 sore dengan transaksi terbanyak terjadi di jam 12 siang yang menjadikan jam 12 siang menjadi *golden hour*.

Perilaku Pembelian Pelanggan Berdasarkan Hari



Gambar 12.

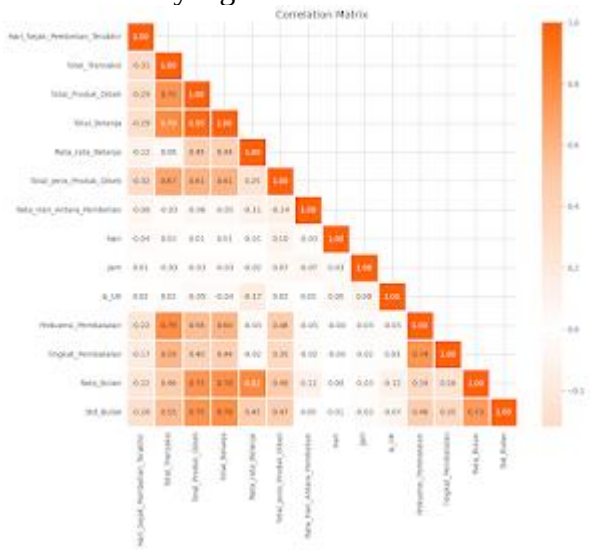
Pada grafik diatas dapat dilihat bahwa Lebih banyak pelanggan yang cenderung melakukan pembelian di hari kamis. Sedangkan pada hari Sabtu, sangat sedikit jumlah pelanggan yang melakukan pembelian pada hari tersebut. Sehingga Sabtu menjadi hari yang paling tidak disukai bagi pelanggan untuk berbelanja atau melakukan transaksi

2.5 Transformasi Data

Transformasi data adalah suatu tahapan atau cara yang digunakan untuk merubah skala suatu data ke dalam bentuk lain sehingga data yang dimiliki mempunyai distribusi yang sesuai dan diharapkan.

2.5.1 Correlation Matrix

Pada tahapan correlation data matrix di bawah ini bertujuan untuk mengetahui hubungan dan keeratan antar variabelnya. Dapat dilihat dari hasil yang didapatkan pada correlation matrix ini berdasarkan parameter dan skala matrik sebagai pengukur banyak dari setiap parameter yang ada. Keterhubungan antar variabel akan dilihat dari hasil yang didapatkan berdasarkan variabel parameter dan skala yang dimiliki.



Gambar 13.

Banyak fitur yang menunjukkan korelasi yang sangat tinggi satu sama lain, yang berarti adanya kemungkinan multikolinearitas. Hal ini dapat ditangani dengan mereduksi kompleksitas data dengan mengurangi dimensinya melalui metode seperti *Principal Component Analysis* (PCA), yang dapat difokuskan pada varian-varian utama dalam data.

2.5.2 Scaling

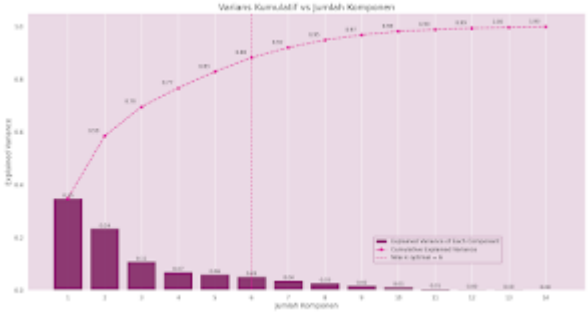
Scaling data merupakan suatu proses yang digunakan untuk melakukan perubahan pada skala data agar memiliki rata-rata nol dan variansi satu. Dalam hal ini untuk data 5 teratas. Hal ini termasuk dalam proses preprocessing yang berperan juga sebagai salah satu hal yang membantu untuk membuat numerik pada data sehingga data memiliki rentang nilai atau skala yang sama.
Pada dasarnya scaling data dapat dilakukan dengan cara melakukan

normalisasi dengan mengubah nilai-nilai sehingga data berada pada rentang 0 sampai dengan 1. Kemudian dilakukan dengan cara standarisasi dengan menskalakan data agar memiliki rata-rata 0 dan standar deviasi 1.

Gambar 14.

2.5.3 PCA

PCA ini merupakan salah satu metode yang digunakan sebagai pereduksi dimensi data. Adanya PCA ini digunakan jika data yang ada memiliki jumlah variabel yang besar dan memiliki korelasi antar variabelnya yang dapat dilihat dari matrix correlation diatas. Selain itu, PCA (Principal Component Analysis) adalah suatu metode yang digunakan untuk mengekstraksi fitur penting dari suatu data. PCA bekerja dengan cara mencari variansi terbesar dari data tersebut, kemudian menemukan fitur baru yang saling tidak berkorelasi dengan satu sama lain.
Maka dalam hal ini dapat dilihat pada gambar hasil visualisasi dibawah bahwasanya varians kumulatif vs jumlah komponen akan dilihat hubungannya dan dianalisis sesuai hasil yang didapat. Dalam hal ini nilai K optimal yang digunakan adalah sampai 6.



Gambar 15.

Berdasarkan hasil diatas maka dilakukan pengujian pada PCA dengan optimal nila PCA yaitu sampai dengan 6 dengan menggunakan 14 variabel yang dijadikan sebagai parameter. Maka dapat dilihat hasilnya pada gambar dibawah ini.

Gambar 16.

2.6 Pemodelan Algoritma Clustering

2.6.1 Algoritma K-Means

Algoritma K-Means adalah sebuah metode dalam analisis klaster (clustering) yang digunakan untuk mengelompokkan data menjadi beberapa kelompok (cluster) berdasarkan kesamaan karakteristik tertentu. Tahapan dari algoritma K-Means adalah pertama-tama, angka K dipilih untuk menentukan jumlah cluster yang diinginkan. Angka ini mewakili berapa banyak kelompok yang akan dibentuk oleh algoritma. Selanjutnya, titik-titik awal atau centroid dipilih secara acak. Centroid ini merupakan representasi pusat dari setiap cluster yang akan terbentuk. Setelah itu, setiap titik data dalam dataset ditetapkan ke centroid terdekat, membentuk cluster sesuai dengan jumlah K yang telah ditentukan sebelumnya.

Langkah berikutnya adalah menghitung varians dalam setiap cluster dan menempatkan centroid baru berdasarkan rata-rata dari titik-titik dalam cluster tersebut. Hal ini

dilakukan untuk meningkatkan akurasi representasi pusat cluster. Proses ini diulang dengan menetapkan kembali setiap titik data ke centroid terdekat yang baru. Langkah ini dilakukan berulang kali untuk memastikan bahwa setiap titik data berada dalam cluster yang optimal. Algoritma terus dijalankan, kembali ke langkah tiga, jika terdapat penugasan ulang, yang menunjukkan adanya perubahan dalam pengelompokan. Jika tidak ada perubahan, iterasi akan berhenti yang menandakan bahwa model K-Means telah selesai. Pada titik ini, telah berhasil membentuk cluster-cluster yang merepresentasikan pola atau struktur dalam dataset.

2.6.2 Algoritma K-Means ++

Algoritma K-Means++ memberikan penyempurnaan pada inisialisasi pusat kluster dalam algoritma K-Means standar. Langkah-langkah ini bertujuan untuk mempercepat konvergensi dan menghasilkan solusi yang lebih optimal. Dalam algoritma K-Means++, Pertama-tama sebuah titik data dipilih secara acak dari dataset sebagai pusat kluster pertama. Langkah ini memberikan titik awal untuk proses pengelompokan. Selanjutnya, jarak kuadrat dihitung dari setiap titik data ke kluster yang telah dipilih sebelumnya. Berdasarkan perhitungan ini, probabilitas pemilihan titik data sebagai pusat kluster berikutnya ditentukan. Pemilihan titik data baru sebagai pusat kluster dilakukan dengan mempertimbangkan probabilitas, memberikan kecenderungan lebih besar bagi titik-titik yang memiliki jarak lebih jauh ke kluster yang sudah ada. Langkah ini diulangi hingga jumlah kluster yang diinginkan terpenuhi.

Proses pemilihan pusat kluster berikutnya dengan probabilitas yang lebih tinggi membantu mendistribusikan kluster awal secara merata di seluruh dataset. Setelah inisialisasi menggunakan K-Means++ selesai, iterasi dilanjutkan dengan menggunakan algoritma K-Means standar. Ini melibatkan perhitungan ulang pusat kluster berdasarkan rata-rata titik data dalam setiap kluster, serta penugasan ulang setiap titik data ke kluster yang memiliki pusat terdekat. Proses iteratif ini berlanjut hingga tidak ada perubahan dalam penugasan cluster atau hingga kriteria konvergensi tertentu terpenuhi. Algoritma K-Means++ secara keseluruhan membawa keunggulan dalam memulai proses pengelompokan data, memberikan hasil akhir yang lebih baik dan lebih cepat dalam mencapai konvergensi.

2.6.3 Algoritma DenMune

DenMune bertujuan untuk melakukan pengelompokan dalam ruang dua dimensi dengan mengidentifikasi kelompok-kelompok kompleks berbentuk dan berdensitas sembarang. Tujuannya adalah untuk mengatasi keterbatasan algoritma pengelompokan lainnya saat menghadapi kelompok-kelompok dengan densitas yang bervariasi, bentuk sembarang, dan kelas data yang tidak seimbang. DenMune dirancang untuk secara otomatis mendeteksi, menghapus, dan mengesampingkan noise dari proses pengelompokan. Algoritma ini menggunakan sistem voting di mana titik data berperan sebagai pemilih, dan hanya mereka yang mendapatkan suara tertinggi yang dianggap sebagai konstruktor kelompok. DenMune tidak memerlukan parameter cutoff dan dapat menghasilkan hasil yang kuat.

Langkah pertama yang dilakukan adalah menerapkan Pengurutan Kanonikal, di mana kumpulan titik disusun dengan cermat berdasarkan jumlah K tetangga terdekat yang dimiliki yang diurutkan secara menurun. Proses ini dilakukan untuk memberikan kerangka yang lebih terstruktur terhadap formasi kluster dalam dataset. Dilanjutkan dengan penghapusan Noise, di mana titik-titik noise tipe-1 dan tipe-2 diidentifikasi dan dihapus. Langkah ini sangat krusial untuk membersihkan dataset dari potensi gangguan yang mungkin memengaruhi akurasi dan keandalan hasil akhirnya.

Setelah berhasil menghilangkan noise, tahap berikutnya adalah Konstruksi dan Propagasi Skeleton. Pada langkah ini, titik-titik yang tersisa dibedakan menjadi dua kategori utama, yaitu *dense points* (*seeds*) dan *low-dense points* (*non-seeds*). Hanya *seeds points* yang diambil untuk membentuk skeleton dari kelompok-kelompok target. Pendekatan ini membantu menitikberatkan perhatian pada titik-titik yang memegang peran penting dalam membentuk struktur kluster. Langkah terakhir adalah menggabungkan *weak points*. *Weak points* yang tetap bertahan setelah proses sebelumnya digabungkan satu per satu dengan kelompok yang memiliki jumlah MNN-seeds terbanyak. Langkah ini bertujuan untuk menyatukan dan memperkuat kluster yang memiliki karakteristik serupa, meningkatkan kohesi antara titik-titik data dalam kluster tersebut.

2.6.4 Algoritma K-Medoids

Algoritma K-Medoids merupakan suatu metode clustering yang berfokus pada penentuan pusat kluster dengan pendekatan yang berbeda dari K-Means. Pada awalnya, K titik data dipilih secara acak sebagai medoid awal dari setiap kluster. Selanjutnya, setiap titik data ditempatkan ke dalam kluster berdasarkan medoid terdekat, yang ditentukan melalui perhitungan jarak menggunakan suatu metrik, seperti jarak Euclidean. Langkah berikutnya melibatkan evaluasi cost dengan menghitung total jarak atau biaya antara setiap titik data dalam kluster dengan medoidnya. Medoid baru kemudian dipilih untuk setiap kluster dengan cara mencoba ganti medoid lama dengan titik data lain dan memilih titik data yang memberikan cost paling rendah. Proses ini diulang secara iteratif hingga tidak ada perubahan medoid atau cost yang signifikan, menandakan konvergensi kluster dan selesainya proses clustering. Keuntungan K-Medoids terletak pada kestabilannya terhadap pencilan, karena pusat kluster merupakan titik data aktual. Meskipun membutuhkan komputasi yang lebih intensif daripada K-Means, K-Medoids menjadi pilihan yang baik dalam situasi di mana keandalan terhadap pencilan lebih diutamakan.

BanditPAM++ merupakan algoritma yang dirancang untuk meningkatkan efisiensi komputasional dari masalah pengelompokan k-medoids sambil tetap mempertahankan hasil pengelompokan yang sama. Tujuan dari BanditPAM++ adalah untuk mempercepat algoritma BanditPAM dengan memanfaatkan dua perbaikan algoritma. Algoritma ini menggunakan struktur BanditPAM untuk mengulang informasi pengelompokan dalam setiap iterasi dan juga memanfaatkan struktur tambahan untuk mengulang informasi di seluruh iterasi yang berbeda. Hal ini memungkinkan BanditPAM++ menjadi jauh lebih cepat daripada BanditPAM, membuatnya lebih cocok untuk dataset yang besar. Algoritma

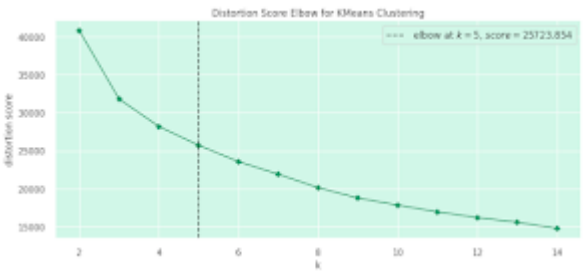
ini menghasilkan solusi pengelompokan yang sama seperti BanditPAM namun dengan waktu eksekusi yang lebih cepat secara signifikan**.**

HASIL DAN PEMBAHASAN

3.1 Algoritma K-Means

3.1.1 Score Elbow Method

Visualisasi menentukan jumlah cluster yang paling sesuai. Grafik Elbow Plot menampilkan nilai inertia (distorsi) terhadap berbagai nilai k (jumlah cluster). Tujuan utamanya adalah untuk menemukan titik di mana penurunan nilai inertia menjadi lebih lambat, sering kali terbentuk seperti siku (elbow). Melalui Gambar visualisasi dibawah menunjukkan jumlah cluster yang optimal pada k=5 dengan score 25723.854, di mana penambahan cluster selanjutnya memberikan sedikit peningkatan signifikan dalam memperbaiki pembagian data ke dalam cluster.

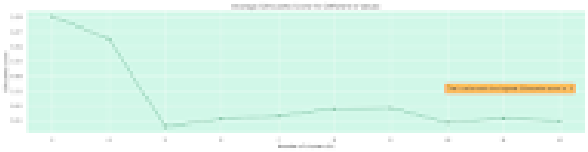


Gambar 17.

3.1.2 Silhouette Score

Pada data yang diberikan dengan tujuan menemukan jumlah cluster yang optimal. Fungsi ini menghasilkan visualisasi yang terdiri dari dua plot utama. Plot pertama menampilkan skor Silhouette rata-rata untuk berbagai nilai k (jumlah cluster) dengan nilai K Value = 3 yang ditentukan sebelumnya dalam rentang tertentu. Pada plot ini, dilakukan iterasi untuk setiap nilai k, di mana model K Means diterapkan pada data dan skor Silhouette dihitung untuk setiap klaster. Selain menampilkan skor Silhouette, plot tersebut juga menandai nilai k yang memberikan skor Silhouette tertinggi sebagai nilai optimal.

Sementara itu, plot kedua menampilkan visualisasi Silhouette untuk masing-masing nilai k dalam rentang yang telah ditentukan sebelumnya. Di setiap subplot, distribusi Silhouette score untuk setiap sampel dalam klaster ditampilkan



Gambar 18.

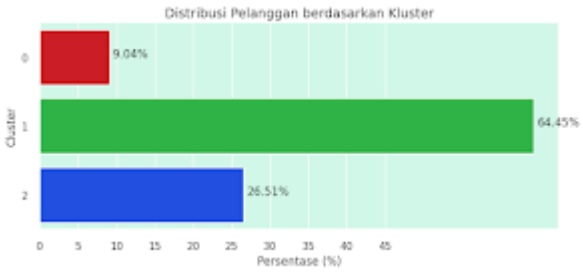
3.1.3 Distribusi Pelanggan berdasarkan Cluster

Hasil persentase pelanggan yang termasuk ke dalam setiap kluster yang telah dibuat sebelumnya. Ini dilakukan dengan menghitung jumlah pelanggan dalam setiap kluster lalu mengkonversi nilai tersebut menjadi persentase dari total jumlah pelanggan. Data ini diurutkan berdasarkan nomor kluster untuk memudahkan interpretasi. Serta dilakukan visualisasi dengan menggunakan diagram batang horizontal. Setiap batang mewakili persentase pelanggan dalam setiap kluster.

Pada setiap batang, ditambahkan label yang menunjukkan persentase

pelanggan yang ada dalam kluster tersebut. terlihat pada Gambar bahwa Cluster 1 mendominasi dengan persentase 64.45% dan Distribusi pelanggan dengan persentase terendah terdapat dalam cluster 0 dengan persentase 9.04%.

Hasil tersebut memberikan pemahaman yang lebih baik tentang proporsi pelanggan di setiap kluster, membantu dalam pengambilan keputusan dan strategi yang sesuai untuk setiap kelompok pelanggan.



Gambar 19.

3.1.4 Matriks Evaluasi

Hasil analisis klasterisasi yang telah dilakukan, terdapat 3806 observasi pelanggan yang telah diproses. Evaluasi menggunakan beberapa metrik penting menunjukkan gambaran tentang seberapa baik pembentukan kluster telah dilakukan. Nilai Silhouette Score yang diperoleh sebesar 0.280 mengindikasikan bahwa sebagian besar sampel data cocok dengan kluster yang dimilikinya, namun terdapat kemungkinan adanya beberapa sampel yang mungkin terletak di kluster yang salah atau memiliki tumpang tindih antar kluster. Sementara itu, nilai Calinski Harabasz Score sebesar 1264.45 menunjukkan bahwa terdapat sebaran yang cukup baik antar kluster dengan variasi yang terlihat di dalam kluster. Nilai Davies Bouldin Score sebesar 1.203 juga menunjukkan adanya pemisahan yang cukup baik antara kluster-kelompok, di mana setiap kluster memiliki kemiripan yang relatif rendah dengan kluster lainnya. Meskipun masih ada ruang untuk peningkatan, hasil evaluasi ini memberikan gambaran yang positif tentang pembentukan kluster dan memberikan dasar yang kuat untuk melanjutkan analisis lebih lanjut terkait karakteristik yang mungkin ada di dalam setiap kluster.

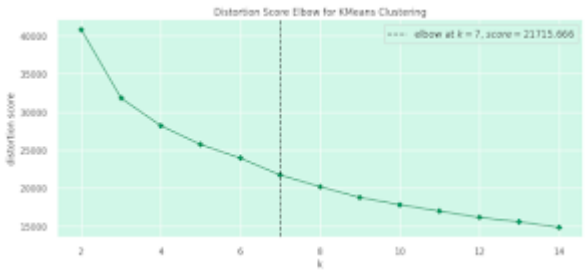
Matriks	Nilai
Jumlah Observasi	3806
Silhouette Score	0.2801908042683165
Calinski Harabasz Score	1264.4500801536456
Davies Bouldin Score	1.2033874268817522

3.2 Algoritma K-Means ++

3.2.1 Score Elbow Method

Distortion Score Elbow K=7
Menentukan jumlah kluster yang optimal untuk data yang sedang diproses dengan menggunakan metode Elbow pada algoritma K Means. Dengan rentang nilai k dari 2 hingga 15, visualisasi Elbow Plot menunjukkan kurva yang menurun secara bertahap seiring dengan peningkatan nilai k. Pada titik di mana

penurunan nilai inertia (distorsi) tidak lagi signifikan atau mencapai suatu titik di mana tambahan kluster tidak memberikan penurunan yang substansial, disebut sebagai siku (elbow) dalam kurva tersebut. Dalam konteks ini, setelah menganalisis Elbow Plot, nilai k=7 dengan score 21715,666 dipilih sebagai jumlah kluster yang optimal. Pada titik ini, terlihat adanya perubahan yang signifikan dalam penurunan nilai inertia yang mulai melandai, menunjukkan bahwa penambahan kluster setelah k=7 tidak memberikan penurunan yang substansial dalam distorsi data.

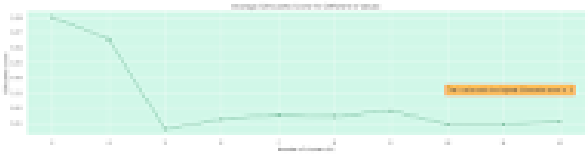


Gambar 20.

Analisis klasterisasi dengan k=7 memungkinkan untuk melakukan segmentasi yang lebih terperinci terhadap data, membedakan kelompok-kelompok yang mungkin memiliki karakteristik, kebutuhan, atau perilaku yang berbeda di antara satu sama lain. Hal ini dapat digunakan untuk mengarahkan strategi pemasaran yang lebih terfokus, personalisasi layanan, atau pengambilan keputusan yang lebih sesuai dengan preferensi masing-masing segmen pelanggan.

3.2.2 Silhouette Score

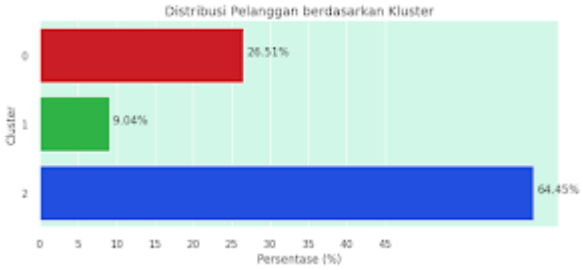
Hasil evaluasi seberapa baik setiap sampel data cocok dengan kluster yang dibentuk dan membantu menentukan jumlah kluster yang optimal. Terdapat plot Silhouette Score yang menampilkan skor Silhouette rata-rata untuk setiap nilai k dalam rentang yang ditentukan (start_k hingga stop_k). berdasarkan visualisasi Plot ini didapatkan nilai k = 3 yang memberikan skor Silhouette tertinggi yaitu 0,28 merupakan pemisahan kluster yang paling baik.



Gambar 21.

3.2.3 Distribusi Pelanggan berdasarkan Cluster

Hasil persentase pelanggan yang terdapat dalam setiap kluster, dengan menggunakan nilai kluster yang disimpan dalam kolom 'cluster' dari data_pelanggan_pca. Berdasarkan Visualisasi distribusi pelanggan tiap Cluster yang ditunjukkan dengan diagram batang. Didapatkan hasil presentase terbesar pelanggan pada Cluster 2 dengan 64,45% dan persentase terendah pada Cluster 1 dengan 9,04%. Hasil tersebut memungkinan mengidentifikasi kluster-kelompok yang mungkin memiliki jumlah pelanggan yang signifikan atau memerlukan perhatian lebih dalam analisis dan strategi pemasaran.



Gambar 22.

3.2.4 Matriks Evaluasi

Hasil analisis klasterisasi yang telah dilakukan, terdapat 3806 observasi pelanggan yang telah diproses. Evaluasi menggunakan beberapa metrik penting menunjukkan gambaran tentang seberapa baik pembentukan kluster telah dilakukan. Nilai Silhouette Score yang diperoleh sebesar 0.280 mengindikasikan bahwa sebagian besar sampel data cocok dengan kluster yang dimilikinya, namun terdapat kemungkinan adanya beberapa sampel yang mungkin terletak di kluster yang salah atau memiliki tumpang tindih antar kluster. Sementara itu, nilai Calinski Harabasz Score sebesar 1264.45 menunjukkan bahwa terdapat sebaran yang cukup baik antar kluster dengan variasi yang terlihat di dalam kluster. Nilai Davies Bouldin Score sebesar 1.203 juga menunjukkan adanya pemisahan yang cukup baik antara kluster-kelompok, di mana setiap kluster memiliki kemiripan yang relatif rendah dengan kluster lainnya. Meskipun masih ada ruang untuk peningkatan, hasil evaluasi ini memberikan gambaran yang positif tentang pembentukan kluster dan memberikan dasar yang kuat untuk melanjutkan analisis lebih lanjut terkait karakteristik yang mungkin ada di dalam setiap kluster.

Matriks	Nilai
Jumlah Observasi	3806
Silhouette Score	0.2801908042683165
Calinski Harabasz Score	1264.4500801536458
Davies Bouldin Score	1.2033874268817522

3.3 Algoritma DenMune

3.3.1 Score Elbow Method

Hasil jumlah kluster yang optimal berdasarkan skor elbow yang dihasilkan. Pada data yang direpresentasikan dalam enam komponen utama (PC1 hingga PC6), setiap pelanggan diproyeksikan ke dalam dimensi yang lebih rendah untuk memfasilitasi proses klasterisasi. Proses menggunakan metode Elbow dalam Denmune Score dilakukan dengan memanfaatkan informasi dari hasil proyeksi pelanggan ke dalam ruang dimensi rendah. Tujuan utamanya adalah untuk menentukan jumlah kluster yang paling optimal berdasarkan elbow point pada kurva Denmune Score. Kurva ini menggambarkan nilai skor yang dihasilkan oleh metrik evaluasi kualitas klaster (Denmune Score) terhadap berbagai kemungkinan jumlah kluster.

Customer ID	PC1	PC2	PC3	PC4	PC5	PC6
12346.0	-1.967151	2.505273	0.653179	-0.196425	-0.018934	0.414643
12347.0	0.737112	-3.320089	-2.320371	-0.104518	0.633046	-0.348125
12348.0	-1.536860	2.517628	-0.076759	-0.578971	0.506549	-0.306079
12349.0	2.210485	-0.114483	-2.656730	1.639909	-1.334167	-0.513711
12351.0	-1.245337	2.519496	-0.429840	-0.628509	1.150464	-0.552750

Gambar 23.

3.3.2 Silhouette Score

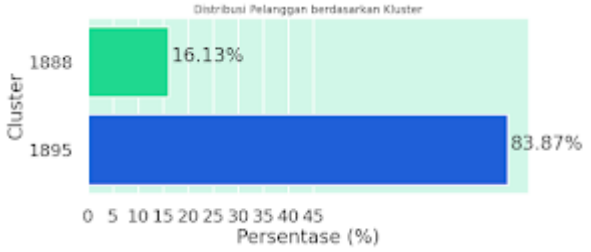
Setiap pelanggan telah dikelompokkan ke dalam suatu kluster tertentu (ditandai dalam kolom 'cluster'). Silhouette Score dihitung untuk setiap sampel data, dan hasilnya memberikan informasi tentang seberapa baik setiap pelanggan cocok dengan kluster tempat ia ditempatkan. Rentang nilai Silhouette Score berkisar dari -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa sampel tersebut ditempatkan dengan benar dalam clusternya dengan adanya pemisahan yang baik antar kluster.

Customer ID	PC1	PC2	PC3	PC4	PC5	PC6	cluster
12346.0	-1.967151	2.505273	0.653179	-0.196425	-0.018934	0.414643	1895
12347.0	0.737112	-3.320089	-2.320371	-0.104518	0.633046	-0.348125	1888
12348.0	-1.536860	2.517628	-0.076759	-0.578971	0.506549	-0.306079	1895
12349.0	2.210485	-0.114483	-2.656730	1.639909	-1.334167	-0.513711	1895
12351.0	-1.245337	2.519496	-0.429840	-0.628509	1.150464	-0.552750	1895

Gambar 24.

3.3.3 Distribusi Pelanggan berdasarkan Cluster

Hasil persentase dari jumlah pelanggan yang terdapat dalam setiap kluster yang dihasilkan dari metode Denmune pada data pelanggan yang direpresentasikan dalam dimensi yang lebih rendah. Didapatkan persentase tertinggi pada cluster 1895 dengan 83.87% dan terendah pada Cluster 1888 dengan 16.13%. hasil tersebut bertujuan mengidentifikasi kluster-kelompok yang mungkin memiliki jumlah pelanggan yang signifikan atau memerlukan perhatian lebih dalam analisis dan strategi pemasaran.



Gambar 25.

3.3.4 Matriks Evaluasi

Terdapat empat hasil metrik evaluasi yang memberikan gambaran tentang kualitas dari klasterisasi yang dilakukan. jumlah observasi atau jumlah sampel yang terdapat dalam dataset adalah sebanyak 3806. Metrik ini memberikan informasi dasar tentang ukuran keseluruhan dari dataset yang digunakan dalam analisis klasterisasi.Terdapat tiga metrik evaluasi yang menggambarkan kualitas dari klasterisasi yang telah dilakukan. Silhouette Score, dengan nilai sebesar 0.2495, memberikan gambaran tentang seberapa baik sampel data cocok dengan kluster yang dimilikinya dibandingkan dengan kluster lainnya. Rentang nilai dari -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan pemisahan kluster yang lebih baik.

Selanjutnya, Calinski Harabasz Score dengan nilai sebesar 719.13 memberikan gambaran tentang seberapa baik kluster-kelompok telah dibentuk dengan mempertimbangkan seberapa besar varians antara kluster dibandingkan dengan varians di dalam kluster. Nilai yang lebih tinggi

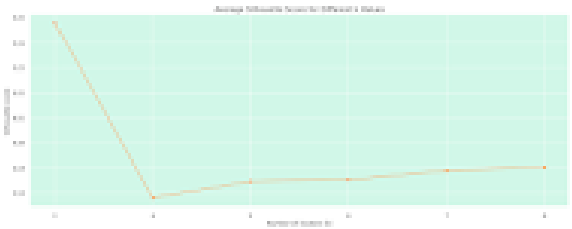
menunjukkan kluster yang lebih terdefinisi dengan baik. Terakhir, Davies Bouldin Score dengan nilai sebesar 1.2364 memberikan evaluasi tentang kemiripan rata-rata antara setiap kluster dengan kluster yang paling mirip dengannya. Nilai yang lebih rendah menunjukkan pemisahan kluster yang lebih baik.

Matriks	Nilai
Jumlah Observasi	3806
Silhouette Score	0.24950565873869757
Calinski Harabasz Score	719.12741971592
Davies Bouldin Score	1.2363903658726652

3.4 Algoritma K-Medoids

3.4.1 Silhoutte Score

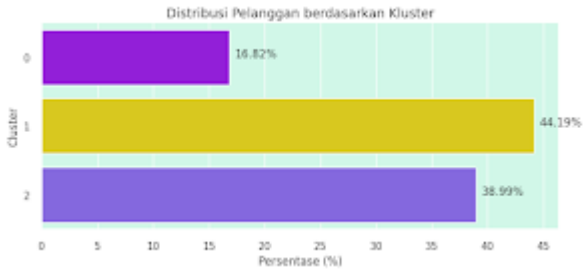
Hasil analisis Silhouette Score untuk berbagai nilai k dalam proses klasterisasi pada data. Didapatkan nilai K tertinggi pada silhoutte adalah cluster k = 3. Pendekatan ini digunakan untuk mengevaluasi seberapa baik setiap sampel data cocok dengan kluster tempatnya berada. Analisis dapat memperoleh informasi penting tentang jumlah klaster yang paling sesuai atau optimal untuk data yang diberikan.



Gambar 26.

3.4.2 Distribusi Pelanggan berdasarkan Cluster

Setelah proses klasterisasi menggunakan algoritma KMedoids. Hasil menghitung persentase dari jumlah pelanggan yang termasuk dalam masing-masing kluster yang dihasilkan. Persentase tersebut disajikan dalam bentuk visualisasi menggunakan diagram batang horizontal. Cluster 1 memiliki Distribusi pelanggan terbanyak dengan persentase 44.19% dan Cluster 0 dengan distribusi pelanggan terkecil dengan jumlah 16.82%. analisis distribusi pelanggan berdasarkan kluster ini dapat memberikan informasi yang berguna dalam memahami karakteristik dan pola-pola yang terdapat di dalam kelompok-kelompok tersebut.



Gambar 27. Distribusi Pelanggan Berdasarkan Cluster

3.4.3 Matriks Evaluasi

Hasil evaluasi matriks klasterisasi data dengan metode K-Medoids terdapat jumlah observasi atau jumlah sampel dalam dataset yang diamati adalah sebanyak 3806. Selanjutnya, nilai Silhouette Score yang

mencapai 0.2477 menggambarkan seberapa baik setiap sampel data sesuai dengan klaster tempatnya ditempatkan dibandingkan dengan klaster lainnya. Rentang nilai Silhouette Score adalah dari -1 hingga 1, dengan nilai yang lebih tinggi menandakan pemisahan klaster yang lebih baik.

Selain itu, metrik Calinski Harabasz Score dengan nilai 1187.1347 memberikan indikasi tentang seberapa jelasnya pemisahan antar klaster dengan mempertimbangkan varian antara klaster dan varian di dalam klaster. Semakin tinggi nilainya, semakin baik klaster-kelompoknya terdefinisikan. Pada metrik Davies Bouldin Score dengan nilai 1.2987 menggambarkan tingkat kemiripan antara setiap klaster dengan klaster lainnya. Semakin rendah nilai Davies Bouldin Score, semakin baik pula pemisahan antar klaster-kelompoknya. Melalui nilai-nilai metrik tersebut, dapat disimpulkan bahwa klasterisasi yang dilakukan pada data menunjukkan hasil yang cukup baik.

Matriks	Nilai
Jumlah Observasi	3806
Silhouette Score	0.24773735406648617
Calinski Harabasz Score	1187.1347675582294
Davies Bouldin Score	1.2987840047433403

KESIMPULAN

Penggunaan algoritma clustering pada big data pelanggan e-commerce tidak hanya memungkinkan perusahaan untuk memberikan layanan yang lebih baik kepada pelanggan, tetapi juga untuk mengoptimalkan pemanfaatan sumber daya mereka. Dalam konteks ekonomi mikro, hal ini berkaitan dengan konsep efisiensi, perilaku konsumen, serta dinamika permintaan dan penawaran di pasar.

DAFTAR PUSTAKA

1. Meewan I, Zhang X, Roy S, Ballatore C, O'Donoghue A, Schooley R, Abagyan R. Discovery of New Inhibitors of Hepatitis C Virus NS3/4A Protease and Its D168A Mutant. ACS Omega. 2019 Okt:4(16):16999–17008. doi:10.1021/acsomega.9b02491 ←

2. Prasaja B, Harahap Y, Lusthom W, Setiawan EC, Ginting MB, Hardiyanti, et al. A bioequivalence study of two tamsulosin sustained-release tablets in Indonesian healthy volunteers. European Journal of Drug Metabolism and Pharmacokinetics. 2011;36(2):109-13. ← Naskah/jurnal dengan lebih dari enam penulis
3. Nakanishi S, Abe M, Yamamoto S, Murai M, Miyoshi H. Bis-THF motif of acetogenin binds to the third matrix-side loop of ND1 subunit in mitochondrial NADH-ubiquinone oxidoreductase. Biochimica et Biophysica Acta. 2011 Sep;1807(9):1170-6. doi: 10.1016/j.bbabio.2011.05.012. ← Naskah/jurnal dengan nomor DOI.
4. Alexander RG. Considerations in creating a beautiful smile. In: Romano R, editor. The art of the smile. London: Quintessence Publishing; 2005. p. 187-210. ← bagian buku
5. Mason J. Concepts in dental public health. Philadelphia: Lippincott Williams & Wilkins; 2005. ← buku dengan nama penulis
6. Norman IJ, Redfern SJ, editors. Mental health care for elderly people. New York: Churchhill Livingstone; 1996. ← Buku dengan nama editor
7. A guide for women with early breast cancer. Sydney: National Breast Cancer; 2003. ← Buku terbitan organisasi, tanpa

nama penulis
8. Kay JG. Intracellular cytokine trafficking and phagocytosis in macrophages [PhD thesis]. St Lucia, Qld: University of Queensland; 2007. ← Thesis/disertasi
9. Canada. Environmental Health Directorate. Radiation protection in dentistry: recommended safety procedures for the use of dental x-ray equipment. Safety Code 30. Ottawa: Ministry of Health; 2000. ← Dokumen pemerintah

LAMPIRAN

Code:https://colab.research.google.com/drive/1fHrQlIxpKSBrNQAFeZHVh1lul7aemBp3?usp=sharing