

Analisis Perilaku Pelanggan E-Commerce melalui Segmentasi Menggunakan K-Means dan Pyspark

Analysis of E-Commerce Customer Behavior through Segmentation Using K-Means and Pyspark

¹Habibie Ed Dien*, ² Annisa Aulia Nadhila, ^NArya Admaja

^{1,2,3}Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang
Jl. Soekarno Hatta No. 9 Kota Malang, Indonesia

*e-mail: habibie@polinema.ac.id

(received: ?, revised: ?, accepted: ? diisi oleh editor)

Abstrak

Dalam era perdagangan secara digital yang berkembang pesat, pemahaman perilaku pelanggan menjadi kunci untuk kesuksesan E-Commerce. Segmentasi pelanggan adalah strategi penting yang membagi pelanggan menjadi kelompok dengan karakteristik, perilaku, atau kebutuhan yang berbeda. Penelitian ini menganalisis perilaku pelanggan E-Commerce melalui segmentasi menggunakan algoritma K-Means dan Pyspark. Metode K-Means digunakan untuk membentuk kelompok pelanggan berdasarkan perilaku pembelian mereka. Hasil analisis menunjukkan beberapa kelompok pelanggan dengan perilaku pembelian yang berbeda, yang dapat digunakan sebagai data pendukung untuk memahami perilaku pelanggan dan menentukan strategi pemasaran yang efektif dan efisien bagi perusahaan E-Commerce. Dalam rangka mengelola hubungan dengan pelanggan dan meningkatkan profitabilitas, penting untuk memahami kebutuhan dan preferensi pelanggan dalam tiap segmen. Dalam konteks ini, penggunaan metode clustering dengan algoritma K-Means membantu menentukan segmentasi pelanggan yang akurat berdasarkan riwayat transaksi. Penelitian ini juga mengaplikasikan model LRFM (Length, Recency, Frequency, dan Monetary) untuk mengenali perilaku dan loyalitas pelanggan.

Kata kunci: Segmentasi pelanggan, K-Means, Pyspark

Abstract

In the rapidly developing digital commerce era, understanding customer behaviour is the key to e-commerce success. Customer segmentation is an essential strategy that divides customers into groups with different characteristics, behaviours or needs. This study analyzes the behaviour of E-Commerce customers through segmentation using the K-Means and Pyspark algorithms. The K-Means method forms customer groups based on their purchasing behaviour. The results of the analysis reveal distinct customer groups with varying purchasing patterns. This data can aid in comprehending customer behaviour and devising successful marketing strategies for E-Commerce businesses. To enhance customer relationships and increase profits, it is crucial to understand the needs and preferences of customers in each segment. In this context, using the clustering method with the K-Means algorithm helps determine accurate customer segmentation based on transaction history. This study also applies the LRFM model (Length, Recency, Frequency, and Monetary) to identify customer behaviour and loyalty.

Keywords: Customer segmentation, K-Means, Pyspark

1 Pendahuluan (or Introduction)

Dalam beberapa tahun terakhir, kemajuan internet telah membawa peningkatan signifikan dalam aktivitas online, termasuk belanja online. Sebagai hasilnya, perusahaan yang beroperasi di sektor ini berhasil mencapai keuntungan yang tinggi. Dengan kesuksesan e-commerce, banyak toko online saling bersaing untuk menawarkan berbagai macam produk dan memberikan pelanggan lebih banyak pilihan.

Namun, jika penjual tidak memiliki pemahaman yang baik tentang minat pembelian calon pelanggan, penawaran yang berlebihan dari berbagai koleksi produk dapat meningkatkan risiko kelebihan stok dan menentukan cara yang tepat dalam mengelola setiap kelompok pelanggan guna meningkatkan keuntungan bagi bisnis. Segmentasi pelanggan membantu perusahaan dalam mengidentifikasi pelanggan yang menguntungkan dan memenuhi kebutuhan mereka melalui penyediaan layanan dan produk yang optimal. Oleh karena itu, segmentasi pelanggan membantu sistem e-commerce dalam mempromosikan perkembangan yang tepat kepada pelanggan yang tepat, sehingga dapat meningkatkan keuntungan. Dalam penelitian ini, perhatian difokuskan pada faktor perilaku pelanggan. Oleh karena itu, algoritma pengelompokan (*clustering*) digunakan untuk menganalisis pola perilaku pembelian dalam sistem e-commerce. *Clustering* bertujuan untuk mencari kesamaan di antara kelompok eksperimental dan memaksimalkan perbedaan antara kelompok-kelompok tersebut.

Dalam penelitian ini, digunakan algoritma pembelajaran untuk memproses data yang dikumpulkan dan mengelompokkan pelanggan, dengan menggunakan metode pengelompokan K-Means. Metode pengelompokan K-Means diimplementasikan untuk mengatasi masalah pengelompokan (*clustering*).

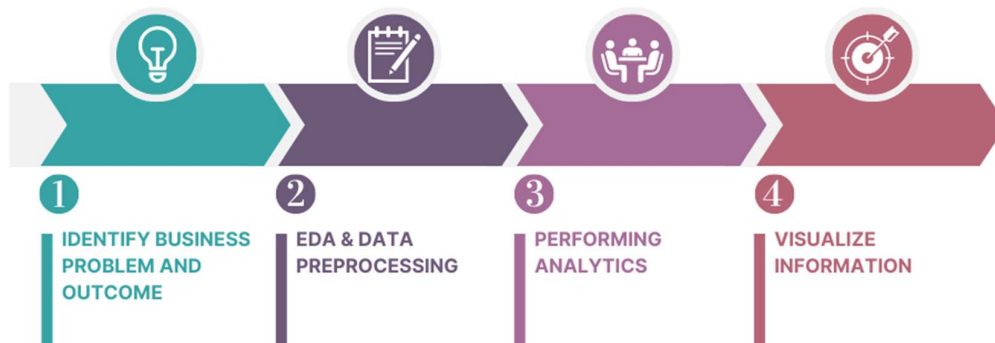
2 Tinjauan Literatur (or Literature Review)

Penelitian terdahulu terkait segmentasi pelanggan dengan metode *clustering* K-Means berdasarkan model RFM seperti penelitian (Adiana, Soesanti and Permanasari, 2018) Aplikasi model RFM dan teknik *clustering* khususnya algoritma K-Means, dapat membantu mengidentifikasi pelanggan dengan nilai tinggi dan berpotensi memberi keuntungan serta dapat membantu dalam merancang strategi pemasaran yang tepat untuk ketiga cluster yang terbentuk.

Penelitian (Harani, Prianto and Nugraha, 2020) hasil segmentasi data pelanggan menggunakan algoritma K-Means dengan nilai $K=3$ telah berhasil mengelompokkan data pelanggan menjadi tiga segmen. Analisis customer profiling pada masing-masing segmen memungkinkan untuk mengidentifikasi karakteristik setiap segmen dengan jelas. Dari simulasi ke-3 dengan presentasi data train 50% dan data test 50% dimana customer profiling cluster 0 dengan kategori pelanggan yang memberikan keuntungan terbesar bagi perusahaan, cluster 1 pelanggan dengan kategori yang tanpa disadari memiliki potensi besar dalam memberikan keuntungan bagi perusahaan, cluster 2 dengan kategori pelanggan yang memberikan keuntungan lebih sedikit dari pada biaya untuk memberikan pelayanan.

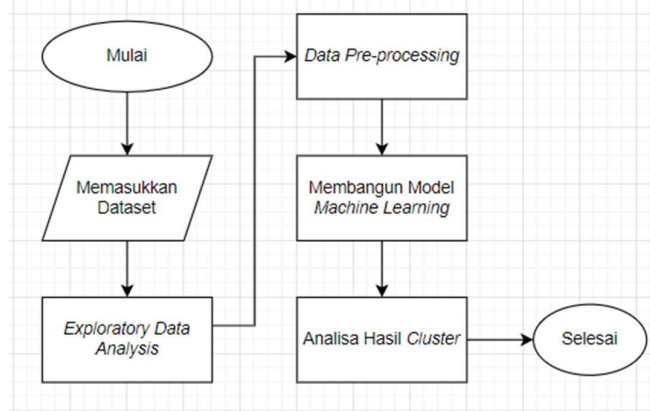
3 Metode Penelitian (or Research Method)

Dalam penerapan big data terdapat beberapa proses yang menjadi alur atau daur hidup proyek, daur hidup ini menjelaskan mengenai runtutan pengolahan data mulai dari mengidentifikasi permasalahan bisnis, mencari data yang diperlukan, pra pengolahan data seperti eksplorasi data (EDA) dan data *pre-processing*, langkah analisa, dan visualisasi hasil analisa. Adapun dalam penggunaannya kami tidak menggunakan keseluruhan proses daur hidup akan tetapi hanya mengambil beberapa tahapan saja sesuai dengan Gambar 1 di bawah ini.



Gambar 1. daur hidup analisa data

Metode penelitian yang digunakan untuk menghasilkan segmentasi pelanggan *e-commerce* digambarkan dalam bentuk *flowchart* yang terdapat pada Gambar 2 berikut :



Gambar 2. *flowchart Tahapan Penelitian*

1. Memasukkan Dataset

Dataset yang digunakan untuk pengujian penelitian ini merupakan data transaksi pelanggan *Online Retail* yang bersumber dari Online Retail. (2015). UCI Machine Learning Repository. <https://doi.org/10.24432/C5BW33>. Data ini terdiri dari 541909 baris data dengan 8 kolom yang masing – masing merupakan InvoiceNo, StockCode, Deskripsi, Kuantitas, Tanggal Faktur, UnitPrice, CustomerID, dan Negara.

2. *Exploratory Data Analysis*

EDA (*Exploratory Data Analysis*) merupakan pendekatan dalam analisis data yang bertujuan untuk secara mendalam mengeksplorasi dan memahami data sebelum dilakukan analisis lebih lanjut. Dalam EDA, dilakukan eksplorasi terhadap struktur data, pola, dan hubungan antar variabel yang terdapat dalam dataset. Pendekatan ini melibatkan penggunaan visualisasi data, perhitungan statistik deskriptif, serta teknik analisis lainnya guna mengidentifikasi anomali, pola menarik, dan informasi penting yang terdapat dalam dataset tersebut. Melalui EDA, dapat diperoleh wawasan awal yang berguna dalam memahami data, menguji hipotesis awal, serta mempersiapkan data sebelum digunakan dalam model atau analisis yang lebih kompleks.

3. *Data Pre-processing*

Data preprocessing merujuk pada serangkaian langkah yang dilakukan sebelum melakukan analisis atau pemodelan data, dengan tujuan membersihkan, mengubah, dan menyesuaikan data mentah. Tujuan dari proses ini adalah untuk mempersiapkan data agar siap digunakan dalam analisis, serta mengurangi kemungkinan terjadinya kesalahan atau bias akibat data yang tidak terstruktur, tidak lengkap, atau tidak konsisten.

Dalam penelitian ini, terdapat beberapa proses data preprocessing yang dilakukan, antara lain:

- i. Transformasi data: Melakukan transformasi data seperti normalisasi untuk mengubah skala data ke dalam rentang tertentu, standarisasi untuk mengubah distribusi data menjadi distribusi normal, atau encoding untuk mengubah data kategorikal menjadi bentuk numerikal.
- ii. Seleksi fitur: Memilih subset fitur yang paling relevan dan informatif untuk analisis atau pemodelan, dengan menghilangkan fitur yang tidak signifikan atau memiliki korelasi tinggi.
- iii. Rekayasa fitur: Menciptakan fitur baru yang memberikan informasi tambahan atau meningkatkan pemahaman tentang data, misalnya dengan menggabungkan beberapa fitur atau melakukan ekstraksi fitur dari data mentah.

Proses data preprocessing ini penting dalam memastikan data yang digunakan dalam analisis atau pemodelan lebih akurat, konsisten, dan dapat diandalkan.

4. Membangun Model Machine Learning

Setelah semua analisis dan persiapan data telah diselesaikan, langkah selanjutnya adalah membangun model pengelompokan K-Means.

Model akan dibuat menggunakan API pembelajaran mesin PySpark.

i. Penentuan Jumlah Cluster yang Digunakan

Ketika model pengelompokan K-Means dibangun, jumlah kelompok atau cluster yang diinginkan perlu ditentukan terlebih dahulu. Misalnya, jika kita memilih tiga cluster, maka akan terdapat tiga segmen pelanggan. Metode ini melibatkan menjalankan algoritma K-Means dengan berbagai jumlah cluster dan memvisualisasikan hasil model untuk setiap cluster. Plot hasilnya akan memiliki titik belok yang menyerupai siku, dan jumlah cluster dipilih berdasarkan titik tersebut.

ii. Pembangunan Model Pengelompokan K-Means

Kurvanya terlihat seperti siku. Dalam plot di atas, siku berada pada $k=3$ (yaitu, Jumlah jarak kuadrat tiba-tiba turun), menunjukkan k optimal untuk kumpulan data ini adalah 3.

iii. Membuat Prediksi

5. Analisa Hasil Cluster

Menganalisa hasil data dari masing – masing *cluster*.

4 Hasil dan Pembahasan

A. Hasil

1. Memasukkan Dataset

Dataset online retailer dapat delapan atribut yaitu InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, dan Country. Empat atribut terpilih sesuai model RFM yaitu Quantity, InvoiceDate, UnitPrice, dan CustomerID.

2. Exploratory Data Analysis

Sebelum melakukan pengolahan data lebih lanjut, dilakukan suatu proses *Exploratory Data Analysis* untuk menganalisa data yang ada pada *Dataset Online Retail*. Dalam prosesnya dapat diketahui bahwa *Dataset Online Retail* memiliki informasi sebagai berikut :

- Jumlah baris yang ada sejumlah 541909
- Jumlah data yang unik ada sejumlah 4373
- Negara yang paling banyak melakukan pembelian adalah “United Kingdom” dengan total pembelian sebanyak 3951
- Penjualan pertama dilakukan pada 2010-12-01 08:26:00
- Penjualan terakhir dilakukan pada 2011-12-09 12:50:00

3. Data Pre-processing

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	date
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	2010-12-01 08:26:00
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010-12-01 08:26:00
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	2010-12-01 08:26:00
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010-12-01 08:26:00
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010-12-01 08:26:00

only showing top 5 rows

Dari dataset yang telah disajikan, kami perlu melakukan segmentasi pelanggan berdasarkan pola pembelian masing-masing pengguna. Format variabel dalam dataset ini tidak dapat digunakan secara langsung dalam model segmentasi pelanggan. Secara individual, fitur-fitur

ini tidak memberikan banyak informasi tentang perilaku pembelian pelanggan. Oleh karena itu, dengan menggunakan variabel yang tersedia untuk dilakukan pengolahan untuk menghasilkan tiga fitur baru yang informatif, yaitu kebaruan, frekuensi, dan nilai moneter (RFM).

Konsep RFM umumnya digunakan dalam bidang pemasaran untuk mengevaluasi nilai pelanggan berdasarkan faktor-faktor berikut:

- A. Keterkinian (*Recency*): Seberapa baru pelanggan melakukan pembelian terakhir?
- B. Frekuensi (*Frequency*): Seberapa sering mereka melakukan pembelian?
- C. Nilai Moneter (*Monetary Value*): Berapa rata-rata jumlah uang yang mereka habiskan saat melakukan pembelian?

A. *Recency*

Untuk menghitung nilai *recency*, yaitu tanggal dan waktu terakhir pembelian dilakukan di platform. Hal Ini dapat dicapai melalui dua tahap:

- i. Skor *recency* untuk setiap pelanggan akan ditentukan dengan mengurangi setiap tanggal dalam dataset dengan tanggal paling awal yang tercatat. Dalam proses ini, dapat terindikasi berapa lama pembelian dari tanggal awal yang tercatat dalam dataset. Skor 0 akan diberikan kepada mereka yang terlihat melakukan pembelian pada tanggal paling awal, menunjukkan *recency* terendah.
- ii. Pemilihan pembelian terbaru akan dilakukan dengan memilih waktu ketika pelanggan terakhir kali terlihat melakukan pembelian. Dalam hal ini, waktu tersebut mencerminkan kapan pembelian terakhir dilakukan oleh pelanggan.

B. *Frequency*

Perhitungan frekuensi nilai akan dilakukan untuk mengukur seberapa sering sesuatu dibeli oleh pelanggan. Untuk melakukan hal ini, pengelompokan akan dilakukan berdasarkan setiap ID pelanggan dan jumlah barang yang dibeli oleh mereka akan dihitung.

C. *Monetary Value*

Perhitungan nilai moneter merupakan jumlah total yang dibelanjakan oleh setiap pelanggan dalam dataset. Hal Ini dapat dicapai melalui dua tahap:

- i. Jumlah total yang dihabiskan dalam setiap pembelian ditemukan: Setiap customerID dilengkapi dengan variabel yang disebut "Quantity" dan "UnitPrice" untuk satu pembelian: Untuk mendapatkan jumlah total yang dibelanjakan oleh setiap pelanggan dalam satu kali pembelian, perlu dilakukan perkalian "Quantity" dengan "UnitPrice".
- ii. Jumlah total yang dihabiskan oleh setiap pelanggan ditemukan: Untuk mengetahui jumlah total yang dibelanjakan oleh setiap pelanggan secara keseluruhan, dilakukan pengelompokan berdasarkan kolom CustomerID dan dilakukan penjumlahan jumlah total yang dibelanjakan. Gabungan dari data ini dengan semua variabel lainnya.

Sebelum model segmentasi pelanggan dibangun, perlu dilakukan standardisasi terhadap kerangka data untuk memastikan bahwa skala semua variabel memiliki keseragaman. Dalam penerapannya dilakukan langkah – langkah sebagai berikut :

- i. Sebuah objek VectorAssembler bernama "assemble" dibuat, yang menggunakan tiga kolom input (*Recency*, *Frequency*, dan *Monetary Value*) dan menghasilkan kolom tunggal yang disebut "fitur".

- ii. Tiga kolom masukan digabungkan menjadi satu vektor yang akan digunakan sebagai masukan untuk algoritma pembelajaran mesin.
- iii. Transformasi *assemble* diterapkan pada *dataset* "finaldf" menggunakan metode transformasi, dan hasilnya disimpan dalam "*assembled_data*".
- iv. Dibuat objek *StandardScaler* yang dinamakan "scale", yang menggunakan kolom fitur dari "*assembled_data*" dan menghasilkan kolom baru yang disebut *standardized*.
- v. Kolom "fitur" dilakukan penskalaan sehingga setiap fitur memiliki rata-rata 0 dan standar deviasi 1.
- vi. Transformasi "scale" diterapkan pada "*assembled_data*" menggunakan metode *fit* dan *transform*, hasilnya disimpan dalam "*data_scale_output*".

4. Membangun *Model Machine Learning*

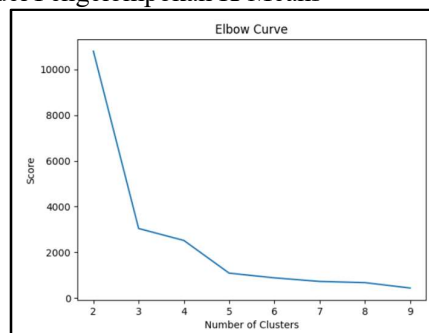
Setelah semua analisis dan persiapan data telah diselesaikan, langkah selanjutnya adalah membangun model pengelompokan K-Means.

Model akan dibuat menggunakan API pembelajaran mesin PySpark.

- i. **Penentuan Jumlah Cluster yang Digunakan**
Ketika membangun model pengelompokan K-Means, tahap awal adalah menentukan jumlah kelompok atau cluster yang akan dihasilkan oleh algoritme. Misalnya, jika menggunakan tiga cluster, maka akan terdapat tiga segmen pelanggan.

Metode yang umum digunakan untuk menentukan jumlah cluster dalam K-Means disebut *elbow method*. Metode ini melibatkan algoritme K-Means dengan berbagai jumlah cluster dan memvisualisasikan hasilnya untuk setiap cluster. Plot yang dihasilkan akan menunjukkan titik-titik tikungan yang membentuk pola seperti siku, dan kita akan memilih jumlah cluster pada titik tersebut.

- ii. **Pembangunan Model Pengelompokan K-Means**



Kurvanya terlihat seperti siku. Dalam plot di atas, siku berada pada k=3 (yaitu, Jumlah jarak kuadrat tiba-tiba turun), menunjukkan k / cluster optimal untuk kumpulan data ini adalah 3.

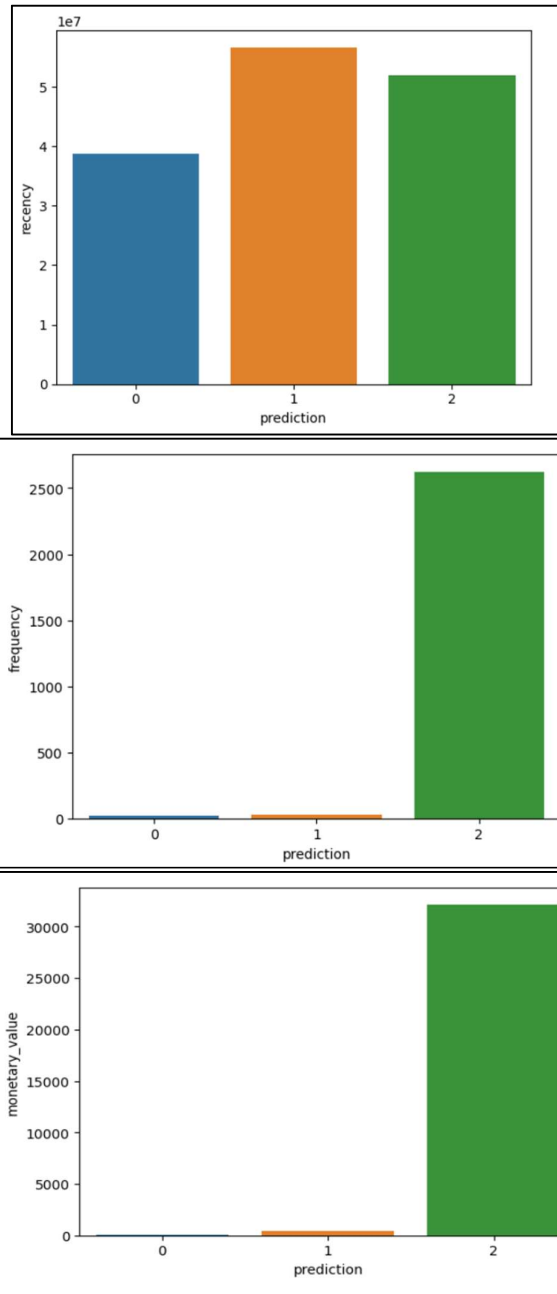
- iii. **Membuat Prediksi**
Manfaatkan model yang telah dibuat untuk mengatribusikan kelompok (cluster) kepada setiap pelanggan dalam *dataset*:

recency	frequency	monetary_value	CustomerID	Features	standardized	prediction
[45885060]	[11]	[184.29000000000002]	[12493.0]	[4.5885067,11.0,184.29000000000002]	[5.254522415034828,0.0381111156188601,0.041256890228470004]	[0]
[43564740]	[11]	[184.29000000000002]	[12493.0]	[4.3564747,11.0,184.29000000000002]	[4.988811234749707,0.0381111156188601,0.041256890228470004]	[0]
[35362500]	[11]	[264.7]	[13973.0]	[3.5362507,11.0,264.7]	[4.049532056199406,0.0381111156188601,0.05925822029862936]	[0]
[44422080]	[33]	[124.14999999999999]	[13533.0]	[4.4422087,33.0,124.14999999999999]	[5.086989427113538,0.11433334685658031,0.027793385001164198]	[0]
[35096040]	[28]	[220.1]	[18114.0]	[3.5096047,28.0,220.1]	[4.019019019675662,0.09701011248437118,0.04927365315147998]	[0]

only showing top 5 rows

Perhatikan bahwa ada kolom "prediksi" dalam *dataset* ini, hal ini yang memberi tahu di *cluster* mana setiap CustomerID berada.

5. Analisa Hasil Cluster



Kesimpulannya, berdasarkan analisis clustering yang dilakukan pada data pelanggan dalam sistem *e-commerce*, dapat diidentifikasi tiga kluster pelanggan dengan karakteristik berbeda.

Kluster 0 menunjukkan bahwa pelanggan dalam segmen ini memiliki nilai *recency* menengah. Adapun nilai *frequency*, dan *monetary_value* yang rendah. Mereka jarang melakukan pembelian di platform dan cenderung menjadi pelanggan potensial yang kurang aktif dalam berbisnis dengan perusahaan *e-commerce*.

Kluster 1 menunjukkan bahwa pengguna dalam kluster ini menunjukkan tingkat *recency* yang tinggi, tetapi belum menghabiskan *monetary_value* yang signifikan di platform. Mereka juga memiliki nilai *frequency* yang rendah. Hal ini mengindikasikan bahwa mereka mungkin merupakan pelanggan baru yang baru saja memulai berbisnis dengan perusahaan.

Kluster 2 menunjukkan bahwa pelanggan dalam segmen ini menunjukkan tingkat *recency*, *frequency* yang tinggi, serta menghabiskan jumlah uang atau *monetary_value* yang signifikan di platform. Hal ini menunjukkan bahwa mereka cenderung membeli barang dengan nilai tinggi atau melakukan pembelian dalam jumlah besar.

Dengan pemahaman tentang karakteristik pelanggan dalam setiap kluster, perusahaan *e-commerce* dapat mengambil langkah-langkah yang sesuai dalam mempertahankan pelanggan yang menguntungkan, meningkatkan keterlibatan pelanggan baru, dan mengoptimalkan strategi penjualan untuk meningkatkan keuntungan bisnis mereka.

i. Pembahasan

Setelah proses segmentasi menggunakan algoritma K-Means, hasil yang diperoleh adalah pembagian pelanggan ke dalam beberapa kluster. Setiap kluster memiliki karakteristik dan perilaku pembelian yang berbeda. Hasil ini dapat digunakan untuk memahami preferensi dan kebutuhan pelanggan di masing-masing kluster.

Pada penelitian ini, dilakukan analisis perilaku pelanggan *e-commerce* melalui segmentasi menggunakan algoritma K-Means dan Pyspark. Hasil analisis clustering menunjukkan adanya tiga kluster pelanggan yang memiliki karakteristik yang berbeda.

- i. Kluster 0 terdiri dari pelanggan yang memiliki nilai *recency* menengah, tetapi memiliki nilai *frequency* dan *monetary_value* yang rendah. Pelanggan dalam kluster ini jarang melakukan pembelian di platform dan cenderung menjadi pelanggan potensial yang kurang aktif dalam berbisnis dengan perusahaan *e-commerce*. Dalam hal ini, perusahaan perlu melakukan strategi pemasaran khusus untuk mendorong pelanggan dalam kluster ini agar lebih aktif dan meningkatkan frekuensi pembelian mereka.
- ii. Kluster 1 terdiri dari pengguna yang menunjukkan tingkat *recency* yang tinggi, namun belum menghabiskan *monetary_value* yang signifikan di platform. Mereka juga memiliki nilai *frequency* yang rendah, menandakan bahwa mereka mungkin merupakan pelanggan baru yang baru saja memulai berbisnis dengan perusahaan. Untuk kluster ini, perusahaan perlu fokus pada upaya retensi pelanggan dan meningkatkan nilai pembelian mereka dengan strategi promosi dan penghargaan.
- iii. Kluster 2 terdiri dari pelanggan yang menunjukkan tingkat *recency* dan *frequency* yang tinggi, serta menghabiskan jumlah uang atau *monetary_value* yang signifikan di platform. Hal ini menunjukkan bahwa mereka cenderung membeli barang dengan nilai tinggi atau melakukan pembelian dalam jumlah besar. Perusahaan dapat memanfaatkan kluster ini sebagai peluang untuk meningkatkan penjualan dengan menawarkan produk premium atau memberikan layanan yang lebih personal kepada pelanggan dalam kluster ini.

Dengan menganalisis perilaku pelanggan melalui segmentasi menggunakan K-Means dan Pyspark, perusahaan *e-commerce* dapat lebih memahami karakteristik pelanggan mereka secara lebih mendalam. Informasi ini dapat digunakan untuk mengembangkan strategi pemasaran yang lebih efektif, meningkatkan retensi pelanggan, dan meningkatkan pendapatan perusahaan.

5 Kesimpulan (or Conclusion)

Penelitian ini telah membahas tentang analisis perilaku pelanggan di industri *e-commerce* melalui segmentasi menggunakan algoritma K-Means dan Pyspark. Penelitian ini berhasil mengidentifikasi 3 *cluster* optimal melalui metode Elbow. *cluster* k optimal diperoleh melalui tahapan preprocessing dan

pembuatan model machine learning. Segmentasi pelanggan di bidang e-commerce dilakukan dengan menggunakan algoritma K-Means dan model RFM, yang merupakan salah satu pilihan alternatif dalam mengatasi permasalahan identifikasi kelompok-kelompok pelanggan. Hasil analisis ini memberikan pemahaman yang mendalam tentang preferensi dan kebutuhan pelanggan dalam setiap kluster, sehingga dapat digunakan sebagai dasar untuk meningkatkan strategi pemasaran dan pengalaman pelanggan. Dengan memanfaatkan teknologi dan metode analisis yang tepat, perusahaan *e-commerce* dapat mengoptimalkan operasional dan mencapai keunggulan kompetitif di pasar yang semakin kompetitif.

Referensi (Reference)

- Adiana, B. E., Soesanti, I. and Permanasari, A. E. (2018) 'Analisis Segmentasi Pelanggan Menggunakan Kombinasi Rfm Model Dan Teknik Clustering', *Jurnal Terapan Teknologi Informasi*, 2(1), pp. 23–32. doi: 10.21460/jutei.2018.21.76.
- Harani, N. H., Prianto, C. and Nugraha, F. A. (2020) 'Segmentasi Pelanggan Produk Digital Service Indihome Menggunakan Algoritma K-Means Berbasis Python', *Jurnal Manajemen Informatika (JAMIKA)*, 10(2), pp. 133–146. doi: 10.34010/jamika.v10i2.2683.