PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 910-917
E-ISSN: 3109-8886

Segmentasi Pelanggan Berbasis RFM dan Analisis Asosiasi Produk pada Olist Brazilian E-Commerce Menggunakan FP-Growth

[1]Cyntiya Olyfiyany, [2]Hayyu Risma Ameilya

[123]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Banten

[1]cyntiyaolyfiyany0524@gmail.com, [2]hayyurisma2@gmail.com

Abstract

This study implements a two-stage analytical pipeline on the Olist Brazilian E-Commerce Public Dataset, integrating RFM-based (Recency, Frequency, Monetary) K-Means Clustering with Market Basket Analysis using the FP-Growth algorithm within the CRISP-DM framework. A total of 92,424 unique customers from 96,478 delivered orders were analyzed. K-Means Clustering at K=4 (Silhouette Score = 0.46) produced four behaviorally distinct customer segments: Potential Loyalists (47,963 customers, 39.9% revenue share), At-Risk/Lost (35,455 customers, 29.7%), Loyal Customers (6,272 customers, 24.9%), and Champions (2,734 customers, 5.4%). FP-Growth analysis, applied exclusively to multi-item transactions per segment, revealed that over 91% of Olist transactions are single-item purchases—a structural characteristic that constrains association rule formation. Nevertheless, meaningful rules emerged: the Loyal Customers segment produced a lift of 16.05 for the product pair bebes → cool_stuff, while At-Risk/Lost and Champions segments showed consistent associations between cama_mesa_banho and casa_conforto (lift 4.12–4.30). The integration of both methods proved that K-Means segmentation enriches FP-Growth by enabling context-specific association analysis, producing more actionable cross-selling recommendations than global analysis alone.

Keywords*: customer segmentation, market basket analysis, K-Means clustering, FP-Growth, RFM, e-commerce, CRISP-DM*

Abstrak

Penelitian ini mengimplementasikan pipeline analitik dua tahap pada dataset Olist Brazilian E-Commerce Public, mengintegrasikan K-Means Clustering berbasis RFM (Recency, Frequency, Monetary) dengan Market Basket Analysis menggunakan algoritma FP-Growth dalam kerangka kerja CRISP-DM. Sebanyak 92.424 pelanggan unik dari 96.478 pesanan berstatus delivered dianalisis. K-Means Clustering dengan K=4 (Silhouette Score = 0,46) menghasilkan empat segmen pelanggan yang terbedakan secara perilaku: Potential Loyalists (47.963 pelanggan, kontribusi revenue 39,9%), At-Risk/Lost (35.455 pelanggan, 29,7%), Loyal Customers (6.272 pelanggan, 24,9%), dan Champions (2.734 pelanggan, 5,4%). Analisis FP-Growth yang dijalankan khusus pada transaksi multi-item per segmen mengungkap bahwa lebih dari 91% transaksi Olist merupakan pembelian satu item—karakteristik struktural yang membatasi pembentukan association rules. Meski demikian, rules bermakna tetap ditemukan: segmen Loyal Customers menghasilkan lift sebesar 16,05 untuk pasangan produk bebes → cool_stuff, sementara segmen At-Risk/Lost dan Champions menunjukkan asosiasi konsisten antara cama_mesa_banho dan casa_conforto (lift 4,12–4,30). Integrasi kedua metode membuktikan bahwa segmentasi K-Means memperkaya analisis FP-Growth dengan memungkinkan analisis asosiasi yang kontekstual per segmen, menghasilkan rekomendasi cross-selling yang lebih tepat sasaran dibandingkan analisis global.

Kata Kunci: segmentasi pelanggan, analisis keranjang belanja, K-Means clustering, FP-Growth, RFM, e-commerce, CRISP-DM

Kata Kunci: tuliskan 3-5 kata kunci di sini, pisahkan dengan tanda koma..

A. PENDAHULUAN

Pada tahun 2021, Brasil adalah pasar digital terbesar di Amerika Latin dan salah satu dari sepuluh pasar e-commerce terbesar di dunia. Dataset E-Commerce Olist Brasil benar-benar menunjukkan fase pertumbuhan yang dinamis karena pertumbuhan ini dimulai dari jangka waktu yang jauh lebih singkat pada tahun 2016, tahun di mana penelitian ini dimulai. Dengan pertumbuhannya, platform marketplace menghadapi masalah struktural yang sama: sebagian besar pelanggan baru tidak bertransaksi lagi setelah membeli barang pertama, yang menghasilkan pola pembelian satu kali yang dominan tanpa disadari oleh pengelola platform.

Untuk membuat strategi retensi yang berhasil, sangat penting untuk memahami dengan baik bagaimana

Diterima Redaksi: 10-06-2026 | Selesai Revisi: 22-06-2026 | Diterbitkan Online: 23-06-2026

pelanggan bertindak. Kerangka RFM (Recency, Frequency, and Monetary) telah terbukti memiliki kemampuan untuk mengkuantifikasi perilaku transaksi pelanggan ke dalam tiga dimensi yang penting bagi bisnis. Kerangka ini juga berfungsi sebagai dasar untuk algoritma clustering berbasis jarak seperti K-Means. Namun, segmentasi saja belum cukup; pemahaman tentang produk apa yang dibeli bersama dalam satu transaksi memungkinkan strategi bundling dan cross-selling yang disesuaikan untuk setiap segmen.

Market Basket Analysis (MBA) dengan algoritma FP-Growth adalah cara yang efisien untuk menemukan pola asosiasi antar produk dari data transaksi yang besar. FP-Growth jauh lebih efisien pada dataset bervolume besar karena struktur FP-Treenya memungkinkan penambangan pola tanpa generasi kandidat berulang, berbeda dengan algoritma Apriori yang membutuhkan pembangkitan kandidat itemset secara eksplisit [3]. Dengan menggabungkan K-Means dan FP-Growth ke dalam satu pipeline analitik, aturan yang dibuat dapat spesifik untuk tiap segmen pelanggan dengan karakteristik perilaku yang berbeda, daripada berlaku untuk seluruh pelanggan.

Dataset Olist Brazilian E-Commerce Public yang tersedia secara publik di Kaggle digunakan. Dataset ini mencakup 115.653 pesanan dari September 2016 hingga Oktober 2018, dan terdiri dari lima tabel relasional. Proses analisis melibatkan pemeriksaan 92.424 pelanggan individu setelah pembersihan data menyeluruh dan penghapusan pesanan berstatus pengiriman. Dataset ini menonjol karena dominasi transaksi satu item sebesar lebih dari 91%. Kondisi struktural ini merupakan masalah utama dalam analisis asosiasi dan merupakan temuan penting dalam penelitian ini.

Dua tujuan penelitian saling melengkapi. Pertama, menggunakan K-Means Clustering pada fitur RFM untuk membentuk segmen pelanggan yang signifikan secara bisnis. Ini dilakukan dengan menggunakan Metode Elbow dan Skor Silhouette untuk menentukan jumlah cluster yang ideal. Selanjutnya, menggunakan FP-Growth untuk mengidentifikasi pola asosiasi produk yang unik untuk masing-masing kelompok pelanggan, sehingga rekomendasi cross-selling yang dihasilkan dapat disesuaikan dengan karakteristik masing-masing kelompok pelanggan. Kerangka CRISP-DM digunakan untuk mengatur seluruh proses penelitian

Fokus utama penelitian ini adalah mengintegrasikan secara eksplisit analisis asosiasi produk dan segmentasi pelanggan dalam satu pipeline. Dengan memanfaatkan hasil K-Means untuk menjalankan FP-Growth secara terfokus per segmen, metode ini menghasilkan rekomendasi yang lebih nyata dibandingkan dengan analisis asosiasi global, karena aturan yang ditemukan pada segmen Loyal Customers, misalnya, tidak serta-merta relevan untuk segmen At-Risk/Lost yang memiliki pola penilaian yang berbeda.

.B. METODE

Pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penyebaran adalah enam tahap iteratif dari proses data mining standar industri CRISP-DM. Kerangka ini dipilih karena agnostik terhadap domain dan dapat mengakomodasi siklus analisis yang berulang, yang memungkinkan penyesuaian pada fase sebelumnya dipicu oleh hasil fase sebelumnya.

1. Pengumpulan dan Deskripsi Data

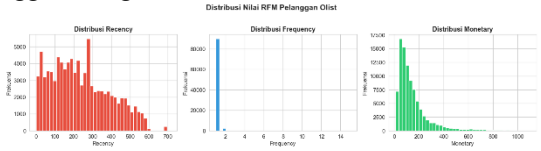
Set data Olist E-Commerce Public Brazilian yang tersedia secara publik melalui platform Kaggle [4] digunakan. Dataset ini terdiri dari lima tabel relasional: tabel pesanan (99.441 baris) berisi informasi tentang status dan waktu pesanan; tabel order_items (112.650 baris) berisi harga dan detail produk per item; tabel pelanggan (99.441 baris) berisi identitas pelanggan; tabel produk (32.951 baris) menyimpan informasi tentang kategori produk; dan tabel pembayaran (103.886 baris) mencatat nilai pembayaran. Data dikumpulkan dari September 2016 hingga Oktober 2018.

2. Pembersihan dan Integrasi Data

Proses integrasi data dimulai dengan menggabungkan lima tabel dengan kunci relasi order_id dan customer_id. Hanya pesanan dengan status pengiriman yang difilter, 96.478 dari 99.441 pesanan, atau 96,9 persen dari total. Dataframe dengan shape diperoleh setelah proses join antar tabel (110.197, 17). Akibat ketidakcocokan product_id antara tabel order_items dan products, ditemukan 1.537 baris dengan nilai kosong pada kolom product_category_name. Nilai kosong ini diberi label "tidak diketahui". Untuk keperluan perhitungan frekuensi, kolom order_purchase_timestamp diubah ke tipe datetime, dan kolom total_price ditambahkan sebagai penjumlahan harga dan harga pengiriman.

3. Pembentukan Fitur RFM

Kerangka RFM digunakan untuk membangun tiga fitur utama dari data transaksi bersih. Tanggal referensi, yang ditetapkan sebagai satu hari setelah tanggal transaksi terakhir dalam dataset, dan tanggal pembelian terakhir pelanggan disebut frekuensi. Nilai frekuensi yang lebih kecil menunjukkan bahwa pelanggan lebih baru bertransaksi. Jumlah total order pelanggan yang unik dihitung sebagai frekuensi. Nilai total per pelanggan dihitung dengan menjumlahkan total harga seluruh pesanan. Setelah perhitungan RFM, nilai di atas persentil ke-99 (P99) dihilangkan, menghilangkan outlier ekstrim pada variabel keuangan, yang menghasilkan 92.424 pelanggan sebagai unit analisis terakhir.



Gambar 1. Distribusi Nilai RFM Pelanggan Olist

4. Pembentukan Fitur RFM

Kerangka RFM digunakan untuk membangun tiga fitur utama dari data transaksi bersih. Tanggal referensi, yang ditetapkan sebagai satu hari setelah tanggal transaksi terakhir dalam dataset, dan tanggal pembelian terakhir pelanggan disebut frekuensi. Nilai frekuensi yang lebih kecil menunjukkan bahwa pelanggan lebih baru bertransaksi. Jumlah total order pelanggan yang unik dihitung sebagai frekuensi. Nilai total per pelanggan dihitung dengan menjumlahkan total harga seluruh pesanan. Setelah perhitungan RFM, nilai di atas persentil ke-99 (P99) dihilangkan, menghilangkan outlier ekstrim pada variabel keuangan, yang menghasilkan 92.424 pelanggan sebagai unit analisis terakhir.

5. Penentuan Jumlah Klaster Optimal

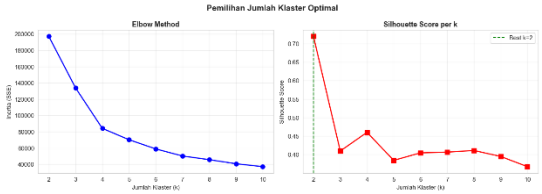
Algoritma K-Means menginisialisasi centroid K dan kemudian mengalokasikan setiap titik data ke centroid terdekat hingga konvergensi. Inisialisasi dilakukan menggunakan metode k-means++ dengan parameter n_init=10. Jumlah cluster yang ideal dipilih menggunakan dua metrik secara bersamaan: Metode Elbow, yang memplot inersia terhadap nilai K dari 2 hingga 10, dan Skor Silhouette, yang mengukur kualitas pemisahan antar klaster. Tabel 1 menunjukkan hasil evaluasi.

Tabel 1. Hasil Evaluasi Elbow Method dan Silhouette Score

K	Inertia	Silhouette Score
2	197.127	0,7194
3	133.855	0,4098
4	84.150	0,4602 ✓
5	70.176	0,3846
6	58.875	0,4049
7	50.178	0,4068
8	45.674	0,4113
9	40.640	0,3951
10	37.110	0,3676

✓ = K yang paling cocok dipilih.

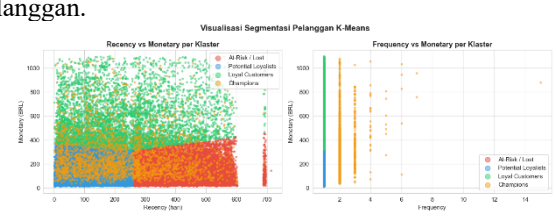
Namun, meskipun K=2 menghasilkan skor Silhouette tertinggi (0,7194), dua klaster tidak memiliki nilai bisnis yang memadai untuk segmentasi pelanggan e-commerce yang kompleks. K=4 dipilih karena menghasilkan skor Silhouette tertinggi di antara nilai K yang menghasilkan segmentasi bermakna (K≥4), yaitu 0,4602. Selain itu, ada penurunan inersia signifikan sebesar 37,1% dari K=3 ke K=4, yang menunjukkan titik siku pada



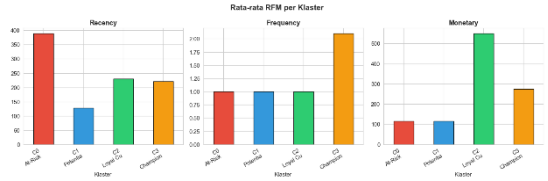
Gambar 2. Jumlah Klaster Optimal

6. Interpretasi dan Pelabelan Segmen

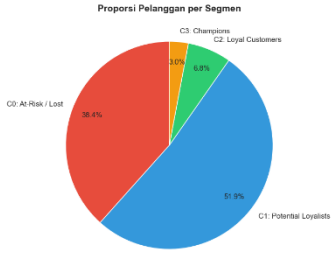
Setelah K-Means dilakukan dengan K=4, setiap klaster dievaluasi berdasarkan nilai rata-rata ketiga fitur RFM. Klaster dengan frekuensi rendah dan nilai moneter dan frekuensi di atas median dilabeli sebagai Champions; klaster dengan frekuensi tinggi tetapi nilai moneter dan frekuensi di bawah median dilabeli sebagai Potential Loyalists; dan klaster dengan frekuensi tinggi tetapi frekuensi di bawah median dilabeli sebagai Potential Loyalists. Label ini kemudian dimasukkan kembali ke dalam data transaksi sebagai fitur segmen untuk setiap pelanggan.



Gambar 3. Segmentasi Pelanggan K-Means



Gambar 4. Rata-Rata RFM Per Klaster



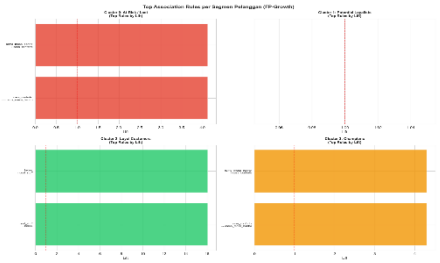
Gambar 5. Proposi Per Segmen

7. Penambangan Pola Asosiasi Produk

Algoritma FP-Growth digunakan untuk melakukan analisis asosiasi produk untuk setiap segmen hasil clustering. FP-Growth dipilih karena mampu menangani dataset transaksi yang sangat besar tanpa memunculkan kandidat itemset secara eksplisit seperti yang dilakukan oleh algoritma Apriori. Algoritma ini membangun struktur FP-Tree yang menampilkan seluruh transaksi dalam bentuk pohon terkompresi dan kemudian secara rekursif menambang pola sering.

Mengingat fakta bahwa lebih dari 91% transaksi pada dataset Olist adalah pembelian satu item, analisis FP-Growth hanya digunakan untuk transaksi multi-item per segmen. Klaster yang memiliki kurang dari dua puluh transaksi multi-item telah dilewati. Formula max(0,005; 5/n), di mana n adalah jumlah transaksi multi-item per klaster, digunakan untuk menentukan nilai minimum support secara adaptif. Karena lift mengukur kekuatan asosiasi relatif terhadap kemungkinan acak, metrik lift

dengan ambang minimum 1,0 digunakan untuk mengekstraksi peraturan asosiasi



Gambar 6. Top Association Rules per Segmen berdasarkan Lift

Gambar 7. Sebaran Support vs Confidence seluruh Rules

8. Penggunaan Skor Silhouette

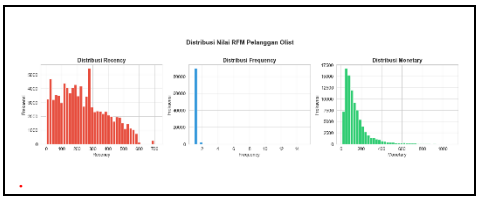
yang mengukur seberapa baik setiap titik data ditempatkan dalam klasternya dibandingkan dengan klaster lain [6]. Nilai di atas 0,5 menunjukkan struktur klaster yang kuat, nilai 0,25 hingga 0,5 adalah moderat, dan nilai di bawah 0,25 menunjukkan struktur klaster yang lemah. Untuk mengevaluasi kualitas peraturan asosiasi, tiga metrik digunakan: support (proporsi transaksi yang memuat kedua item), confidence (probabilitas bersyarat pembelian terkait jika antecedent dibeli), dan lift (rasio keyakinan terhadap keyakinan yang diharapkan jika kedua item independen). Fokus utama analisis dan rekomendasi adalah peraturan dengan lift tinggi dan keyakinan yang tinggi.

C. HASIL DAN PEMBAHASAN

Bagian ini menampilkan hasil dari seluruh tahapan analisis yang telah dilakukan. Hasil ini mencakup eksplorasi distribusi fitur RFM, hasil segmentasi pelanggan dengan K-Means Clustering, profil yang dibentuk untuk setiap segmen, dan hasil pola asosiasi produk yang diperoleh melalui algoritma FP-Growth per segmen. Setiap temuan diikuti dengan diskusi tentang konsekuensi bisnisnya.

1. Distribusi Fitur RFM

Sebelum proses clustering, analisis distribusi ketiga fitur RFM dilakukan untuk memahami karakteristik umum basis pelanggan Olist. Fitur waktu memiliki distribusi yang sangat miring ke kanan (right-skewed), yang menunjukkan bahwa sebagian besar pelanggan memiliki jarak waktu yang cukup panjang sejak transaksi terakhir mereka—sebuah indikasi awal dominasi pelanggan yang hanya bertransaksi satu kali. Fitur frekuensi memiliki distribusi yang sangat miring ke kanan, yang menunjuk Dengan distribusi yang miring ke kanan dan beberapa nilai ekstrim yang ditunjukkan oleh fitur moneter, pemotongan diperlukan pada persentil ke-99 sebelum clustering dapat dilakukan.

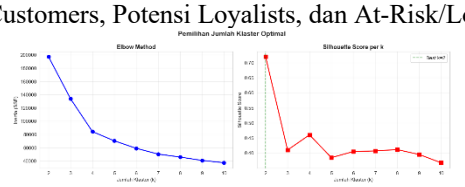


Gambar 8. Distribusi Nilai RFM Pelanggan Olist

2. Menentukan Jumlah Klaster Terbaik

Percobaan kelompokan dilakukan pada rentang K=2 hingga K=10. Menurut Metode Elbow, kurva inersia menunjukkan penurunan yang tajam dari K=2 (197.127) hingga K=4 (84.150), kemudian melandai secara bertahap setelah K=4. Pola ini menunjukkan bahwa penambahan klaster di atas K=4 tidak lagi memberikan peningkatan kualitas pemisahan yang signifikan secara proporsional.

Dari sudut pandang Silhouette Score, nilai tertinggi diperoleh pada K=2 (0,7194), tetapi dua klaster tidak menghasilkan segmentasi yang cukup granular untuk kebutuhan bisnis. Jumlah klaster K=4 dianggap ideal karena menghasilkan nilai Silhouette Score tertinggi di antara nilai K yang signifikan secara bisnis (K lebih dari 4), yaitu 0,4602, yang berada di rentang moderat (0,25–0,5), yang menunjukkan struktur klaster yang baik. Keputusan ini juga sejalan dengan praktik umum dalam literatur RFM untuk membagi pelanggan menjadi 4 kategori: Champions, Loyal Customers, Potensi Loyalists, dan At-Risk/Lost.



Gambar 9. Elbow Method dan Silhouette Score per K

3. Profil dan Karakteristik Segmen Pelanggan

Clustering K-Means dengan K=4 menghasilkan empat segmen pelanggan yang berbeda secara karakteristik RFM. Profil masing-masing segmen disajikan pada Tabel 2.

Tabel 2. Profil Rata-rata RFM per Segmen Pelanggan

Segmen	Pelanggan	Recency (hari)	Frequency	Monetary (BRL)	Revenue Share (%)
Potential Loyalists	47.963	128.23	1.00	115.14	39.90%
At-Risk/Lost	35.455	388.95	1.00	115.79	29.70%
Loyal customers	6.272	230.78	1.00	549.17	24.90%
Champions	2.734	220.99	2.10	274.98	5.40%

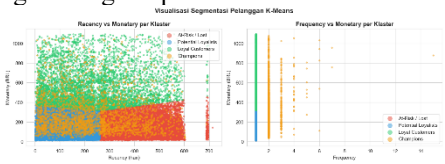
Potential Loyalists merupakan, sebagai segmen terbesar Potential Loyalists memiliki 47.963 pelanggan, yang

merupakan 51,9% dari total, dan menghasilkan pendapatan tertinggi sebesar 39,9%. Segmen ini menonjol karena frekuensi yang relatif rendah (128,23 hari), yang menunjukkan bahwa mereka masih aktif dalam rentang waktu yang relatif singkat, tetapi frekuensi tetap 1,00, yang menunjukkan bahwa sebagian besar dari mereka belum melakukan pembelian ulang. Melalui penerapan strategi retensi yang tepat, segmen ini paling mungkin menjadi pelanggan setia.

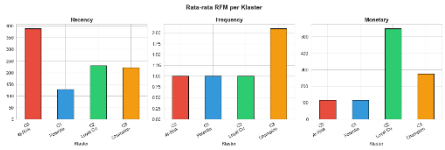
At-Risk/Lost adalah segmentasi memiliki jumlah pelanggan terbesar kedua (35.455 pelanggan), tetapi memiliki frekuensi paling tinggi, 388,95 hari, yang berarti sudah hampir 13 bulan tidak bertransaksi. Pelanggan segmen ini mungkin telah meninggalkan platform Olist karena nilai moneter yang rendah (R$115,79) dan frekuensi yang rendah. Untuk menghidupkan kembali segmen ini, intervensi kampanye win-back diperlukan.

Loyal Cutomer, meskipun hanya 6.272 pelanggan—atau 6,8% dari total pelanggan—mencatat nilai moneter rata-rata R$549,17 per pelanggan, hampir 4,8 kali lipat dari nilai rata-rata Potential Loyalists. Meskipun pelanggannya hanya sedikit, nilai ekonominya sangat besar, dengan kontribusi pendapatan sebesar 24,9% dari total pendapatan. Meskipun frekuensi sebesar 230,78 hari menunjukkan bahwa ada transaksi yang cukup lama, nilai moneter yang tinggi menunjukkan bahwa mereka memiliki nilai yang besar saat berbelanja.

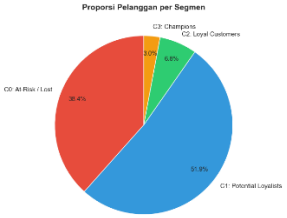
Champions adalah segmen terkecil dengan 2.734 pelanggan (3,0%). Namun, itu adalah satu-satunya segmen yang memiliki frekuensi di atas 1,00 (rata-rata 2,10 transaksi), menunjukkan bahwa ada pembelian berulang. At-Risk dan Potential Loyalists memiliki nilai moneter rata-rata R$274,98. Komponen ini harus diprioritaskan dalam program loyalitas jangka panjang karena merupakan aset paling berharga di platform.



Gambar 10. Sebaran Klaster pada Ruang Fitur RFM



Gambar 11. Rata-rata Nilai RFM per Segmen



Gambar 12. Proporsi Jumlah Pelanggan per Segmen

4. Hasil penambangan Pola Asosiasi Produk

Analisis FP-Growth dilakukan secara terpisah pada transaksi multi-item per segmen. Dari total transaksi yang ada, lebih dari 91% adalah transaksi satu item, yang merupakan karakteristik struktural dataset Olist yang membatasi pembentukan peraturan persatuan secara umum. Dari empat segmen yang ada, hanya tiga yang menghasilkan peraturan persatuan yang signifikan, sementara segmen Potential Loyalists tidak menghasilkan peraturan karena jumlah transaksi multi-item yang tersedia tidak membentuk pola yang cukup konsisten di atas ambang minimum dukungan. Tabel 3 menunjukkan ringkasan peraturan yang telah dibuat.

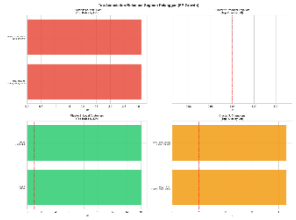
Tabel 3. Association Rules Hasil FP-Growth per Segmen

Segmen	Antecedent	Consequent	Support	Confidence	Lift
At-Risk/Lost	Cama_mesa_banho	Casa_conforto	0.01	0.05	4.12
At-Risk/Lost	-	-	-	-	-
Loyal customers	bebes	Cool_stuff	0.01	0.44	16.05
Loyal Customer	Cool_stuff	bebes	0.01	0.18	16.05
Champions	Cama_mesa_banho	Casa_conforto	0.01	0.03	4.30
Champions	Casa_conforto	Cama_mesa_banho	0.01	1.00	4.30

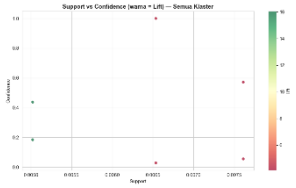
Salah satu hasil yang paling menonjol adalah rule bebes → cool_stuff pada segmen Loyal Customers, yang memiliki lift sebesar 16,05, menunjukkan bahwa pelanggan dalam segmen Loyal Customers yang membeli produk dalam kategori perlengkapan bayi (bebes), memiliki kecenderungan 16 kali lebih besar untuk juga membeli produk dalam kategori cool_stuff daripada pelanggan yang dipilih secara acak. Dengan tingkat kepercayaan 0,44, dapat disimpulkan bahwa 44 persen pelanggan setia yang membeli Bebes juga membeli Cool Stuff dalam transaksi yang sama. Ini adalah hubungan yang sangat kuat dalam data e-commerce yang didominasi oleh transaksi satu item.

Salah satu hasil yang paling signifikan adalah rule bebes → cool_stuff pada segmen Loyal Customers, yang memiliki lift sebesar 16,05, menunjukkan bahwa pelanggan dalam segmen Loyal Customers yang membeli barang dalam kategori perlengkapan bayi (bebes) memiliki kecenderungan 16 kali lebih besar untuk juga membeli barang dalam kategori cool_stuff daripada pelanggan yang dipilih secara acak. Dengan tingkat kepercayaan 0,44, dapat disimpulkan bahwa empat puluh empat persen pelanggan setia Bebes yang membeli Cool Stuff juga melakukan pembelian serupa. Ini adalah korelasi yang sangat kuat dalam data e-commerce yang sebagian besar terfokus pada transaksi satu item.

Tidak ada peraturan di segmen Potential Loyalists karena mereka adalah pelanggan yang masih belajar tentang platform dan memiliki pola pembelian yang belum terbentuk secara konsisten. Pola asosiasi antar kategori belum terbentuk karena pelanggan saat ini cenderung membeli satu kategori produk setiap kali, seperti yang ditunjukkan oleh dominasi transaksi satu item pada segmen ini (91,4%).



Gambar 13. Top Association Rules per Segmen berdasarkan Lift



Gambar 14. Sebaran Support vs Confidence Seluruh Rules

5. Keterkaitan K-Means Clustering dan FP-Growth

Integrasi kedua metode dalam satu pipeline analitik menunjukkan hasil yang lebih kaya. K-Means berfungsi sebagai filter kontekstual yang membagi pelanggan ke dalam kelompok yang memiliki perilaku yang sama, dan FP-Growth dapat menemukan pola asosiasi yang benar-benar unik untuk tiap kelompok. Tanpa segmentasi terlebih dahulu, rule bebes → cool_stuff dengan lift 16,05 yang hanya berlaku untuk Loyal Customers kemungkinan besar akan terselip dalam noise data global atau tidak akan muncul sama sekali karena tidak memiliki dukungan yang cukup untuk pelanggan secara keseluruhan.

Sebaliknya, hasil FP-Growth membantu memahami profil K-Means. Selain menunjukkan nilai moneter yang tinggi, segmen pelanggan setia yang secara RFM dikenal memiliki pola pembelian lintas kategori yang unik, informasi yang tidak dapat diperoleh melalui analisis kelompok. Kedua hasil ini memungkinkan perancangan strategi yang jauh lebih terarah. Misalnya, rekomendasi produk bundling Bebes-Cool_stuff dapat disesuaikan untuk segmen pelanggan yang setia, dan kampanye cross-sell produk rumah tangga dapat dirancang dengan cara yang berbeda untuk segmen Champions (yang memberikan penghargaan atas loyalitas) dan At-Risk/Lost (yang memberikan insentif untuk kembali bertransaksi).

D. PENUTUP

Pipeline analitik dua tahap yang mengintegrasikan K-Means Clustering berbasis RFM dengan Market Basket Analysis menggunakan FP-Growth pada dataset E-Commerce Olist Brasil berhasil dilaksanakan oleh penelitian ini. Hasil analisis telah mengarah pada beberapa kesimpulan berikut.

Simpulan

Pertama, empat segmen pelanggan yang terbedakan secara perilaku dibentuk melalui K-Means Clustering dengan K=4. Skor Silhouette 0,4602 berada pada rentang moderat, yang menunjukkan struktur klaster yang cukup baik. Potential Loyalists (47.963 pelanggan, 39,9% revenue), At-Risk/Lost (35.455 pelanggan, 29,7% revenue), Loyal Customers (6.272 pelanggan, 24,9% revenue), dan Champions (2.734 pelanggan, 5,4% revenue) adalah empat segmen yang membentuk bisnis.

Kedua, setiap segmen memiliki ciri-ciri RFM unik dan dapat digunakan dalam konteks bisnis. Sebagai satu-satunya segmen dengan tingkat frekuensi di atas 1,00 (rata-rata 2,10 transaksi), segmen Champions mencatatkan rata-rata moneter tertinggi (R$549,17) meskipun memiliki pelanggan yang lebih kecil. Ini menunjukkan loyalitas pelanggan yang konsisten terhadap pembelian berulang. Sebuah segmen At-Risk/Lost dengan rentang waktu rata-rata 388,95 hari harus diperhatikan karena memiliki kemungkinan untuk kehilangan hampir 30% dari kontribusi pendapatan.

Ketiga, analisis FP-Growth menunjukkan bahwa dataset Olist memiliki karakteristik struktural yang signifikan: lebih dari 91% transaksi merupakan pembelian satu item (single-item), yang berarti bahwa ada batas terbatas untuk pembentukan peraturan asosiasi. Meskipun demikian, aturan penting masih ada di tiga dari empat bagian. Salah satu hasil yang paling signifikan adalah rule bebes → cool_stuff pada segmen Loyal Customers dengan lift 16,05, yang menunjukkan kecenderungan pembelian bersama 16 kali lebih tinggi daripada acak. Selain itu, ditemukan asosiasi konsisten antara cama_mesa_banho dan casa_conforto pada segmen At-Risk/Lost dan Champions dengan lift 4,12–4,30.

Keempat, ketika K-Means dan FP-Growth diintegrasikan dalam satu pipeline, hasilnya lebih kaya. Ini karena segmentasi berbasis K-Means memungkinkan FP-Growth untuk menemukan pola asosiasi yang spesifik untuk setiap kelompok pelanggan, yang memungkinkan rekomendasi produk yang dihasilkan bersifat kontekstual dan dapat digunakan secara langsung sebagai strategi cross-selling yang dipersonalisasi.

Saran

Pertama, untuk mengatasi dominasi transaksi satu item yang membatasi analisis asosiasi, penelitian lanjutan dapat mempertimbangkan analisis MBA pada tingkat yang lebih mendalam, misalnya dengan menggunakan nama produk individual atau kombinasi kategori dan subkategori. Dengan demikian, kemungkinan pembentukan aturan yang signifikan akan meningkat.

Kedua, penelitian lebih lanjut harus difokuskan pada segmen Potential Loyalists, yang merupakan segmen terbesar (51,9% pelanggan) tetapi belum menghasilkan peraturan asosiasi. Mengingat data mencakup dua tahun periode transaksi, pendekatan analisis sekuensial atau time-series dapat digunakan untuk memahami pola pembelian multi-sesi pelanggan dalam segmen ini.

Ketiga, profil klaster yang lebih kaya secara informasi dapat dihasilkan dengan menambah fitur di luar RFM, seperti kategori produk favorit, metode pembayaran, wilayah geografis, atau rating ulasan. Untuk membandingkan kualitas segmentasi, Anda juga dapat mempertimbangkan metode clustering alternatif seperti Model Mixture Gaussian atau DBSCAN.

Keempat, dari sudut pandang implementasi bisnis, saran strategis yang dihasilkan penelitian ini harus divalidasi melalui uji coba A/B di platform yang nyata. Dalam program win-back campaign, segmen At-Risk/Lost, yang mencakup hampir 30% pendapatan, harus diprioritaskan. Di sisi lain, segmen Loyal Customers dapat menjadi target utama dari strategi bundling yang didasarkan pada hasil FP-Growth.

E. DAFTAR PUSTAKA

[1] eMarketer, "Latin America Ecommerce 2022: Countries to Watch as the Region Recovers," eMarketer Report, New York, USA, 2022.

[2] T. L. Dzulfikar dan A. Adiwijaya, "Customer Segmentation Using K-Means Clustering Based on RFM Model," in Proc. 2019 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2019, pp. 542–547, doi: 10.1109/ICOIACT.2019.8784938.

[3] J. Han, J. Pei, dan Y. Yin, "Mining Frequent Patterns without Candidate Generation," in Proc. ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 2000, pp. 1–12, doi: 10.1145/342009.335372.

[4] A. Olist, "Brazilian E-Commerce Public Dataset by Olist," Kaggle, 2018. [Online]. Tersedia: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. [Diakses: 1 Mei 2025].

[5] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, dan R. Wirth, "CRISP-DM 1.0: Step-by-Step Data Mining Guide," SPSS Inc., Chicago, IL, USA, Technical Report, 2000.

[6] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[7] A. Shabani, M. Sohrabi, dan S. Nik, "RFM Based Customers Segmentation Using Business Intelligence Tools," International Journal of Computer Science and Information Security, vol. 14, no. 7, pp. 1–7, 2016.

[8] R. Agrawal dan R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in Proc. 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, 1994, pp. 487–499.

[9] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 1967, vol. 1, pp. 281–297.

[10] S. Anitha dan M. Patil, "RFM Model for Customer Purchase Behavior Using K-Means Algorithm," Journal of King Saud University – Computer and Information Sciences, vol. 34, no. 5, pp. 1785–1792, 2022, doi: 10.1016/j.jksuci.2019.12.011.

[11] M. Hahsler, C. Buchta, B. Gruen, dan K. Hornik, "arules: Mining Association Rules and Frequent Itemsets," R Package Version 1.7-5, 2022. [Online]. Tersedia: https://CRAN.R-project.org/package=arules.

[12] D. Arthur dan S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), New Orleans, LA, USA, 2007, pp. 1027–1035.

[13] G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules," in Knowledge Discovery in Databases, G. Piatetsky-Shapiro dan W. J. Frawley, Eds. Cambridge, MA: AAAI/MIT Press, 1991, pp. 229–248.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[15] S. Raschka, J. Patterson, dan C. Nolet, "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine

Learning, and Artificial Intelligence," Information, vol. 11, no. 4, p. 193, 2020, doi: 10.3390/info11040193.