

SIMULASI DAN PEMODELAN BIG DATA ANALYTICS DALAM CRM E-COMMERCE INDONESIA

Eisyaniah Desvazulinda¹, Fendi Hidayat², Isdianto³

Sistem Informasi, Fakultas Teknik, Universitas Batam, Jl. Uniba No.5, 29432, Indonesia

*email : e.desvazulinda@univbatam.ac.id¹ , fendihidayat@univbatam.ac.id²

Abstract

Indonesia's e-commerce sector generates customer data volumes that exceed the analytical reach of conventional Customer Relationship Management (CRM) systems, yet few studies model this relationship on transparent, reproducible data. This study develops a simulation and modeling framework for Big Data Analytics (BDA) within CRM decision-making, using Tokopedia as a reference case. A 5,000-record synthetic customer dataset was generated and calibrated in Python against patterns reported in prior literature and national e-commerce statistics, then processed through descriptive statistics, logistic regression, and K-Means clustering. The model identified four customer segments with churn probabilities ranging from 3.7 percent among loyal premium customers to 27.3 percent among discount-driven buyers, with satisfaction score emerging as the strongest predictor at 82 percent classification accuracy. These results indicate that segment-specific retention strategies built on interpretable, simulation-based models can inform CRM decisions even where real transactional data remains inaccessible to external researchers. The study contributes a reproducible simulation and modeling pipeline that links customer segmentation, churn prediction, and Indonesia's Personal Data Protection Law into a single decision framework for CRM practitioners.

Keywords: Big Data Analytics, churn prediction, CRM, customer segmentation, simulation modeling

Abstrak

Pertumbuhan sektor e-commerce Indonesia menghasilkan volume data pelanggan yang melampaui kapasitas analitik sistem Customer Relationship Management (CRM) konvensional, namun masih sedikit penelitian yang memodelkan hubungan tersebut pada data yang transparan dan dapat direproduksi. Penelitian ini menyusun kerangka simulasi dan pemodelan Big Data Analytics (BDA) dalam pengambilan keputusan CRM, dengan Tokopedia sebagai rujukan studi kasus. Dataset sintetis 5.000 pelanggan dibangkitkan dan dikalibrasi dengan Python berdasarkan pola pada literatur terdahulu serta statistik e-commerce nasional, kemudian diproses melalui statistik deskriptif, regresi logistik, dan klasterisasi K-Means. Model yang dihasilkan mengidentifikasi empat segmen pelanggan dengan probabilitas churn mulai dari 3,7 persen pada pelanggan loyal premium hingga 27,3 persen pada pemburu diskon, dengan skor kepuasan sebagai prediktor terkuat pada akurasi klasifikasi 82 persen. Hasil ini menunjukkan bahwa strategi retensi per segmen yang dibangun di atas model simulasi yang interpretatif dapat menjadi dasar pengambilan keputusan CRM sekalipun data transaksi riil tidak dapat diakses peneliti eksternal. Penelitian ini menyumbangkan alur simulasi dan pemodelan yang dapat direproduksi, menghubungkan segmentasi pelanggan, prediksi churn, dan Undang-Undang Pelindungan Data Pribadi ke dalam satu kerangka pengambilan keputusan bagi praktisi CRM.

Kata Kunci: Big Data Analytics, CRM, pemodelan simulasi, prediksi churn, segmentasi pelanggan

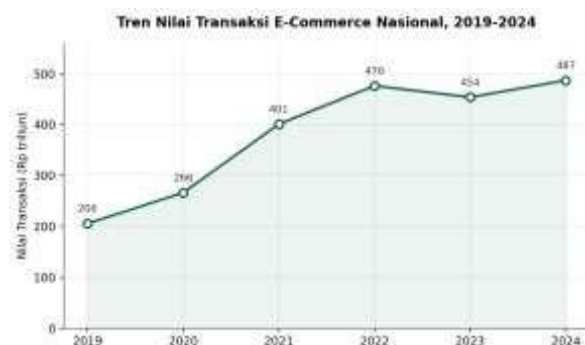
1. Pendahuluan

1.1 Latar Belakang

Transaksi e-commerce nasional tumbuh dari Rp205,5 triliun pada 2019 menjadi Rp487,01 triliun pada 2024, dengan lonjakan tertinggi

terjadi pada 2021 akibat percepatan adopsi belanja daring selama pandemi, sebagaimana ditunjukkan pada Gambar 1. Perlambatan sesaat pada 2023 justru menandakan pasar memasuki fase kematangan, bukan penurunan minat konsumen,

sehingga kompetisi bergeser dari sekadar akuisisi pengguna baru menuju upaya mempertahankan pelanggan yang sudah ada. Pada saat bersamaan, laporan e-Conomy SEA 2025 mencatat nilai transaksi bruto (GMV) e-commerce Indonesia mencapai 71 miliar dolar AS, tumbuh 14 persen secara tahunan (Google, Temasek, & Bain & Company, 2025), sementara struktur pasar semakin terkonsentrasi pada tiga pemain utama setelah konsolidasi Tokopedia dengan TikTok Shop (Momentum Works, 2024). Kombinasi pertumbuhan transaksi yang melambat dan persaingan platform yang menegat menjadikan retensi pelanggan sebagai variabel yang menentukan keberlangsungan bisnis, bukan sekadar indikator pelengkap.



Gambar 1. Tren Nilai Transaksi E-Commerce Nasional 2019-2024 (Sumber: Bank Indonesia, 2025, diolah)

Urgensi ini semakin nyata ketika dilihat dari sisi ekonomi retensi. Literatur pemasaran sejak lama mencatat bahwa biaya menarik pelanggan baru dapat mencapai lima hingga tujuh kali lipat dibandingkan biaya mempertahankan pelanggan lama, sehingga setiap penurunan tingkat churn memberikan dampak berganda terhadap profitabilitas dibandingkan penambahan pengguna baru dalam jumlah yang sama. Konteks ini menjadi lebih relevan setelah konsolidasi Tokopedia dengan TikTok Shop pada awal 2024, yang di satu sisi memperluas basis pengguna melalui integrasi ekosistem konten dan perdagangan, tetapi di sisi lain memunculkan tantangan baru dalam pemantauan data pelanggan lintas platform (Momentum Works, 2024).

Skala data yang dihasilkan dari interaksi, ulasan, dan pola pembelian jutaan pelanggan pada platform sebesar Tokopedia tidak dapat lagi diproses secara memadai melalui pendekatan

CRM konvensional yang mengandalkan data historis dan segmentasi demografis statis. Big Data Analytics (BDA) menawarkan jalan keluar dengan mengolah aliran data menjadi wawasan yang dapat ditindaklanjuti, mencakup prediksi churn, segmentasi berbasis perilaku, dan personalisasi promosi.

Persoalannya, literatur mengenai BDA dan CRM sesungguhnya sudah cukup luas, namun sebagian besar masih berhenti pada tataran konseptual. Del Vecchio dkk. (2022) menegaskan melalui tinjauan literatur terstruktur bahwa BDA mengubah CRM dari sistem reaktif menjadi prediktif, tetapi kajian tersebut bersifat tinjauan pustaka dan tidak menyertakan pengujian model pada data perilaku pelanggan yang sesungguhnya.

Akter dan Wamba (2016) menegaskan hal serupa melalui tinjauan sistematis atas literatur BDA pada e-commerce, tanpa menghadirkan validasi empiris pada dataset tertentu. Khan dkk. (2024) memperluas tinjauan ini dengan menelaah 100 artikel mengikuti pedoman PRISMA dan menyimpulkan bahwa analitik prediktif memungkinkan CRM mengantisipasi perilaku pelanggan, tetapi tinjauan sistematis semacam ini secara desain tidak menghasilkan dataset atau model orisinal yang dapat diuji ulang oleh pembaca.

Ketika penelitian empiris memang tersedia, keterbatasannya bergeser ke persoalan lain. Anitha dan Patil (2022) membuktikan efektivitas kombinasi model RFM dan algoritma K-Means dalam mensegmentasi pelanggan ritel menggunakan data transaksi riil, namun penelitian tersebut berhenti pada tahap segmentasi dan tidak diperluas menjadi prediksi churn maupun kerangka pengambilan keputusan CRM. Arif Lim dan Al-Mudimigh (2018) membahas personalisasi layanan berbasis BDA secara konseptual dan studi kasus, tanpa model kuantitatif yang dapat direplikasi.

Wamba dkk. (2017) memang menghadirkan model kuantitatif yang kuat secara statistik mengenai kapabilitas BDA dan kinerja perusahaan, tetapi unit analisisnya berada pada level organisasi, bukan pada level segmen atau individu pelanggan, sehingga tidak dapat langsung diterjemahkan menjadi keputusan retensi yang operasional. Di Indonesia sendiri, Prakasa dan Subardono (2022) telah menerapkan algoritma clustering pada data

konsumen e-commerce, namun kajian tersebut juga berhenti pada tahap segmentasi tanpa menghubungkannya dengan prediksi churn maupun implikasi kebijakan retensi.

Celah ketiga terletak pada dimensi regulasi data yang jarang dimasukkan sebagai bagian dari kerangka analitik itu sendiri. Roche dan Jamal (2021) menyoroti isu etika dalam big data melalui tinjauan sistematis, tetapi kajian tersebut bersifat umum dan tidak dikaitkan dengan kerangka hukum spesifik suatu negara. Padahal di Indonesia, sejak berlakunya Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi, kepatuhan data telah menjadi prasyarat hukum yang melekat pada setiap tahap pemrosesan data pelanggan bagi penyelenggara sistem elektronik, termasuk platform e-commerce (Republik Indonesia, 2022).

Tidak satu pun dari literatur di atas menggabungkan tiga hal sekaligus, yaitu model yang dapat direproduksi, unit analisis pada level segmen pelanggan, dan pertimbangan regulasi privasi data Indonesia, ke dalam satu kerangka simulasi yang utuh.

Berdasarkan pemetaan tersebut, tiga celah penelitian dapat dirumuskan secara eksplisit. Pertama, celah replikabilitas, karena studi konseptual (Del Vecchio dkk., 2022; Akter & Wamba, 2016; Khan dkk., 2024) tidak menyediakan data maupun kode yang dapat diuji ulang. Kedua, celah kedalaman analisis pada studi empiris yang tersedia (Anitha & Patil, 2022; Prakasa & Subardono, 2022), yang menghentikan analisisnya pada tahap segmentasi tanpa memperluasnya menjadi model prediksi churn yang dapat menjadi dasar keputusan retensi. Ketiga, celah integrasi regulasi, karena kajian tata kelola data yang ada (Roche & Jamal, 2021) belum menyatukan pertimbangan hukum dengan proses pemodelan itu sendiri. Penelitian ini diposisikan sebagai upaya menutup ketiga celah tersebut sekaligus melalui pendekatan simulasi dan pemodelan yang transparan.

Simulasi dipilih bukan sebagai jalan pintas metodologis, melainkan sebagai respons langsung terhadap kendala akses data. Tokopedia sebagai objek studi tidak membuka data transaksi pelanggan secara publik, sehingga peneliti eksternal mana pun menghadapi hambatan yang

sama seperti yang membatasi studi-studi sebelumnya. Dengan membangun dataset simulasi yang parameternya dikalibrasi terhadap pola pada literatur dan statistik transaksi digital nasional, penelitian ini menghasilkan model segmentasi dan prediksi churn yang mekanismenya dapat diperiksa, diuji ulang, dan diadaptasi oleh peneliti maupun praktisi lain, sekaligus tetap merepresentasikan karakteristik pasar e-commerce Indonesia secara wajar. Pendekatan ini menempatkan simulasi dan pemodelan bukan sebagai pengganti data riil, melainkan sebagai jembatan metodologis ketika data riil tidak dapat diakses tanpa melanggar privasi pelanggan.

Berdasarkan uraian di atas, penelitian ini bertujuan untuk: (1) membangun simulasi dan model Big Data Analytics yang dapat mendukung pengambilan keputusan berbasis Customer Relationship Management pada e-commerce Tokopedia, (2) memodelkan segmentasi dan karakteristik perilaku pelanggan sebagai dasar penyusunan strategi CRM yang tepat sasaran, dan (3) menganalisis kontribusi hasil pemodelan tersebut terhadap peningkatan retensi dan kepuasan pelanggan, termasuk implikasinya terhadap tata kelola data. Ketiga tujuan ini saling berkaitan, karena model yang dibangun pada tujuan pertama menjadi dasar segmentasi pada tujuan kedua, yang pada gilirannya diuji relevansi praktisnya pada tujuan ketiga.

2. Tinjauan Pustaka

2.1. Landasan Teori

Penelitian ini bertumpu pada dua kerangka teori yang saling melengkapi. Pertama, model Recency, Frequency, Monetary (RFM) yang menilai pelanggan berdasarkan seberapa baru transaksi terakhir dilakukan, seberapa sering pelanggan bertransaksi, dan seberapa besar nilai uang yang dibelanjakan. Model ini menjadi dasar konseptual bagi variabel frekuensi transaksi dan pengeluaran bulanan yang digunakan dalam proses klusterisasi pada penelitian ini, sejalan dengan penerapannya oleh Anitha dan Patil (2022) pada data ritel.

Kedua, konsep Customer Lifetime Value (CLV) yang memandang pelanggan bukan sebagai transaksi tunggal, melainkan sebagai aliran nilai jangka panjang yang perlu dijaga melalui strategi

retensi. Kedua kerangka ini secara teoretis saling terhubung dengan teori churn prediction, yang berargumen bahwa penurunan frekuensi, nilai transaksi, dan kepuasan merupakan sinyal dini sebelum pelanggan benar-benar berhenti bertransaksi. Ketiga teori ini digunakan sebagai dasar pemilihan variabel dan interpretasi hasil pada bagian analisis, bukan sekadar rujukan normatif.

2.2 Big Data Analytics dan Transformasi CRM

Secara konseptual, CRM dibangun untuk menjalin hubungan jangka panjang yang menguntungkan antara perusahaan dan pelanggan melalui sinergi manusia, proses, dan teknologi. Model CRM konvensional bertumpu pada data historis dan atribut demografis yang statis, sehingga responsnya cenderung tertinggal terhadap perubahan perilaku pelanggan yang berlangsung cepat. BDA mengoreksi keterbatasan ini dengan memungkinkan pengolahan data volume besar, kecepatan tinggi, dan format beragam secara near real-time, mengubah CRM dari arsip pasif menjadi instrumen prediktif.

Akter dan Wamba (2016) menunjukkan bahwa penerapan BDA pada berbagai titik interaksi pelanggan memberi perusahaan visibilitas langsung terhadap dinamika pasar dan preferensi konsumen yang bergeser, sedangkan Del Vecchio dkk. (2022) menempatkan pergeseran ini sebagai perubahan paradigma dalam pemasaran internasional, dari CRM sebagai gudang data menjadi CRM sebagai mesin pengambilan keputusan.

2.3 Segmentasi, Prediksi Churn, dan Perbandingan Pendekatan

Tiga fokus utama penerapan BDA pada e-commerce yang teridentifikasi dari literatur adalah segmentasi pelanggan, prediksi perilaku, dan personalisasi layanan, tetapi setiap studi umumnya hanya menekankan satu dimensi tanpa memodelkan keterkaitan antar-dimensi tersebut secara empiris. Anitha dan Patil (2022) membuktikan bahwa kombinasi RFM dan K-Means menghasilkan segmentasi yang tervalidasi secara statistik melalui koefisien Silhouette pada

data ritel, tetapi tidak memperluasnya menjadi model prediksi churn. Khan dkk. (2024) melalui tinjauan sistematisnya menitikberatkan pada potensi analitik prediktif untuk mendeteksi dini pelanggan berisiko churn, namun tidak menyertakan dataset maupun model orisinal karena sifatnya yang merupakan tinjauan pustaka. Prakasa dan Subardono (2022) sebaliknya menerapkan algoritma clustering pada konsumen e-commerce Indonesia, namun berhenti pada tahap segmentasi tanpa mengaitkannya dengan prediksi churn. Tabel 1 merangkum perbandingan tersebut.

Studi	Jenis Studi	Teknik	Keterbatasan Terkait Penelitian Ini
Del Vecchio dkk. (2022)	Tinjauan literatur teoretis	Transformasi CRM oleh BDA	Konseptual, tanpa model empiris pada data pelanggan
Akter & Wamba (2016)	Tinjauan sistematis	Agenda riset BDA pada e-commerce	Tidak mencakup pemastian model prediktif
Khan dkk. (2024)	Tinjauan sistematis (PRISMA, 100 artikel)	Teknik prediktif dalam CRM	Mengabaikan tren, tanpa dataset atau model orisinal
Anitha & Patil (2022)	Studi empiris	Segmentasi RFM dan K-Means pada data ritel	Tidak mencakup prediksi churn maupun konteks e-commerce Indonesia
Prakasa & Subardono (2022)	Studi empiris (Indonesia)	Klasifikasi konsumen e-commerce	Berfokus di segmentasi, tanpa prediksi churn atau tata kelola data
Penelitian ini	Simulasi dan pemodelan	Segmentasi, prediksi churn, tata kelola data	Data simulasi, bukan dataset riil platform

Tabel 1. Perbandingan Jenis Studi dan Keterbatasan Penelitian Terdahulu

Dari sisi infrastruktur, kerangka kerja terdistribusi seperti Hadoop dan Apache Spark tetap menjadi pilihan dominan untuk analisis data berskala besar. Ahmed dkk. (2020) melalui pengujian benchmark HiBench menunjukkan bahwa Spark dapat mencapai kecepatan hingga empat belas kali lipat dibandingkan Hadoop pada beban kerja tertentu, meskipun performa keduanya tetap bergantung pada ukuran data dan konfigurasi parameter. Ketersediaan infrastruktur seperti ini menegaskan bahwa hambatan penerapan BDA pada CRM e-commerce Indonesia bukan lagi soal teknologi pemrosesan, melainkan soal tata kelola dan kompetensi analitik sumber daya manusia yang menerjemahkan hasil pemrosesan tersebut menjadi keputusan bisnis.

2.4 Tata Kelola

Roche dan Jamal (2021) mengidentifikasi berbagai isu etika dalam lingkungan big data melalui tinjauan sistematis, termasuk bias algoritmik dan risiko penyalahgunaan data pelanggan, namun memposisikan tata kelola sebagai isu etis yang bersifat umum dan belum

dikaitkan dengan kerangka hukum negara tertentu. Di Indonesia, sejak berlakunya UU PDP 2022, kepatuhan data telah menjadi prasyarat hukum yang melekat pada setiap tahap pemrosesan data pelanggan, bukan sekadar praktik etis yang bersifat sukarela (Republik Indonesia, 2022). Wamba dkk. (2017) menunjukkan bahwa kapabilitas analitik berbasis big data berkorelasi dengan kinerja perusahaan melalui peningkatan kapabilitas dinamis organisasi, namun studi tersebut belum menempatkan kepatuhan regulasi sebagai variabel yang memengaruhi keberhasilan kapabilitas tersebut. Penelitian ini memosisikan tata kelola data bukan sebagai catatan penutup, melainkan sebagai salah satu dari tiga pilar keberhasilan integrasi BDA-CRM yang diuji pada bagian hasil.

3. Metode Penelitian

3.1 Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif berbasis simulasi dan pemodelan, diperkuat interpretasi kualitatif-deskriptif atas hasilnya. Pendekatan ini dipilih karena Tokopedia sebagai objek studi tidak membuka akses data transaksi pelanggan secara publik, sehingga simulasi menjadi cara yang sah untuk memodelkan mekanisme BDA-CRM tanpa melanggar batasan privasi data riil, sekaligus menghasilkan alur kerja yang dapat direproduksi oleh peneliti lain.

3.2. Sumber dan Teknik Pengumpulan Data

Dataset dibangun secara terprogram menggunakan Python (NumPy dan Pandas) sebanyak 5.000 baris data pelanggan, dengan parameter distribusi (rata-rata pengeluaran, skor kepuasan, dan probabilitas churn per segmen) dikalibrasi berdasarkan pola yang dilaporkan dalam literatur CRM e-commerce serta statistik transaksi digital nasional. Atribut dataset dirangkum pada Tabel 2. Pendekatan ini konsisten dengan metode simulasi yang lazim digunakan ketika data primer dari platform tertutup tidak dapat diakses peneliti eksternal.

Atribut	Deskripsi
ID Pelanggan	Nomor identitas unik setiap pelanggan
Kelompok Usia	Dikategorikan menjadi 18-23, 24-35, 36-45, dan di atas 45 tahun
Tingkat Rata-rata Bulanan	Frekuensi pembelian pelanggan per bulan
Pembelian Bulanan (DB)	Total rata-rata yang dibelanjakan pelanggan per bulan
Tingkat Keterlibatan	Kategori aktivitas pelanggan: tinggi, sedang, atau rendah
Skor Kepuasan Pelanggan	Nilai 1-10 yang menggambarkan kepuasan terhadap layanan Tokopedia
Status Churn	Menentukan apakah pelanggan berhenti bertransaksi (ya/tidak)

Tabel 2. Atribut Dataset Simulasi Pelanggan

3.3 Teknik Analisis

Metode analisis yang digunakan meliputi:

1. Analisis Deskriptif untuk memetakan tren umum dan karakteristik pelanggan berdasarkan usia, frekuensi transaksi, tingkat keterlibatan, dan pengeluaran, disajikan melalui rata-rata, median, dan distribusi data.
2. Regresi logistik untuk memodelkan probabilitas churn dan mengidentifikasi variabel yang paling berpengaruh terhadapnya, dengan tingkat kepercayaan p kurang dari 0,05.
3. Klasterisasi K-Means untuk memodelkan pengelompokan pelanggan berdasarkan pola transaksi dan pengeluaran, sehingga strategi retensi dapat disusun per segmen, bukan secara seragam.
4. Visualisasi data menggunakan Python (Matplotlib) untuk menyajikan tren, perbandingan, dan korelasi antar-variabel secara konsisten dan dapat direproduksi.

Pemilihan K-Means dan regresi logistik didasarkan pada pertimbangan interpretabilitas model, bukan semata akurasi. Kedua algoritma tersebut menghasilkan parameter yang mudah dijelaskan kepada pengambil keputusan non-teknis, yaitu titik pusat klaster dan koefisien odds ratio, dibandingkan model kotak hitam seperti random forest atau gradient boosting yang lebih sulit diaudit untuk kepentingan kepatuhan terhadap UU PDP. Jumlah klaster ditentukan melalui evaluasi nilai inertia pada rentang dua hingga delapan klaster, dengan empat klaster dipilih karena memberikan penurunan inertia paling signifikan sebelum kurva melandai, pendekatan yang sejalan dengan validasi kualitas klaster yang dilakukan Anitha dan Patil (2022). Model regresi logistik dilatih pada 80 persen data secara acak dan

diuji pada 20 persen sisanya, dengan variabel independen berupa frekuensi transaksi, pengeluaran bulanan, tingkat keterlibatan, dan skor kepuasan.

Seluruh proses, mulai dari pembangkitan data hingga visualisasi akhir, dijalankan melalui skrip Python yang sama sehingga hasil pada bagian berikutnya dapat diverifikasi ulang tanpa bergantung pada perangkat lunak berbayar. Konsistensi alur ini menjadi jawaban langsung terhadap celah replikabilitas yang diuraikan pada bagian latar belakang.

3.4 Kerangka Penelitian

Kerangka penelitian menggambarkan alur integrasi antara pengumpulan data, pemrosesan, analitik, dan pengambilan keputusan dalam sistem CRM, sebagaimana disajikan pada Gambar 2. Umpan balik dari keputusan yang diambil dialirkan kembali ke tahap pengumpulan data, menjadikan proses ini siklikal dan terus diperbarui, bukan proyek pemodelan satu kali jalan.



Gambar 2. Kerangka Kerja Big Data Analytics dalam Pengambilan Keputusan CRM

4. Hasil dan Pembahasan

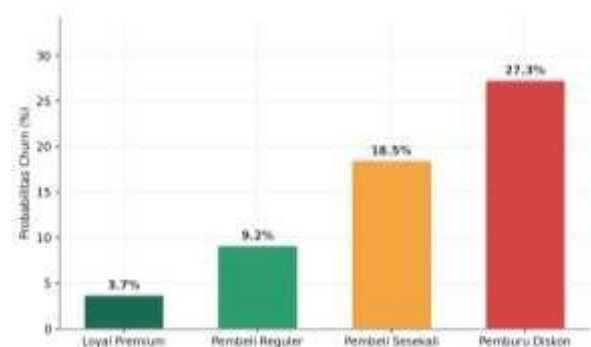
4.1 Segmentasi Pelanggan

Klasterisasi K-Means terhadap 5.000 data simulasi menghasilkan empat segmen pelanggan dengan karakteristik yang berbeda signifikan, dirangkum pada Tabel 3 dan divisualisasikan pada Gambar 3 dan Gambar 4.

Segmen Pelanggan	Pengeluaran Bulanan Rata-rata (IDR)	Skor Kepuasan	Probabilitas Churn	Proporsi Persentase
Loyal Premium	2.394.633	9,1	3,7%	22%
Pembeli Reguler	1.101.681	7,3	9,2%	46%
Pembeli Sesekali	598.583	5,8	18,5%	20%
Pemburu Diskon	340.404	4,9	27,3%	12%

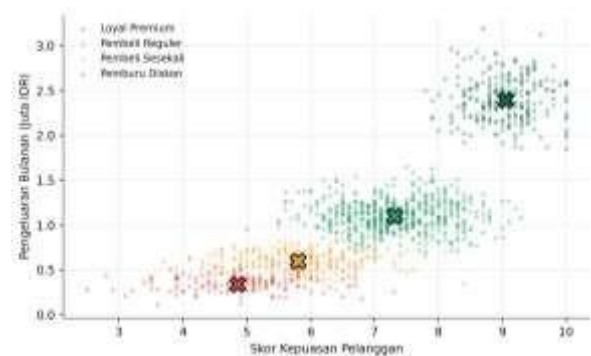
Tabel 3. Segmentasi Pelanggan Hasil Pemodelan Simulasi

Pola yang muncul tidak sekadar mengonfirmasi bahwa pelanggan bernilai tinggi lebih loyal, melainkan menunjukkan hubungan yang hampir linear antara pengeluaran, kepuasan, dan churn, sebagaimana terlihat pada Gambar 3.



Gambar 3. Tingkat Churn per Segmen Pelanggan

Selisih probabilitas churn antara segmen Loyal Premium dan Pemburu Diskon mencapai 23,6 poin persentase. Rentang selisih yang lebar ini konsisten dengan argumen Khan dkk. (2024) bahwa memasukkan variabel kepuasan ke dalam model prediktif memperkuat daya bedanya dibandingkan model yang hanya mengandalkan atribut demografis. Artinya, memasukkan variabel kepuasan dan keterlibatan ke dalam model bukan sekadar meningkatkan akurasi statistik, tetapi mengubah cara perusahaan memprioritaskan anggaran retensi antar-segmen.



Gambar 4. Hubungan Skor Kepuasan dan Pengeluaran Bulanan per Segmen

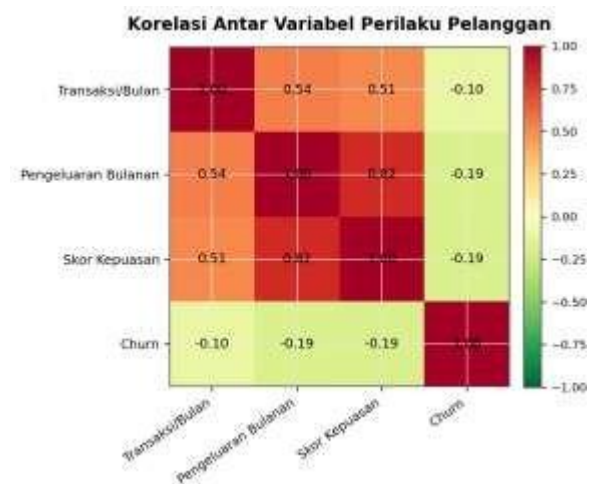
Gambar 4 memperlihatkan pemisahan klaster yang cukup tegas pada tingkat individu pelanggan, bukan hanya pada rata-rata segmen, sejalan dengan validitas klaster yang ditunjukkan Anitha dan Patil (2022) melalui koefisien Silhouette pada konteks data ritel. Hal ini penting karena kritik umum terhadap studi segmentasi, termasuk yang dilakukan Prakasa dan Subardono (2022), adalah minimnya bukti bahwa klaster yang dihasilkan benar-benar terpisah pada level data mentah dan bukan artefak dari agregasi statistik semata. Pola pemisahan yang konsisten pada Gambar 4 juga menguatkan validitas teori RFM yang dipaparkan pada bagian tinjauan pustaka, karena dua dimensi yang membentuk sumbu grafik tersebut, yaitu kepuasan sebagai proksi loyalitas dan pengeluaran sebagai komponen monetary, terbukti bergerak searah dan membentuk empat kelompok berbeda tanpa memerlukan variabel tambahan.

Menariknya, sebaran data pada segmen Pembeli Reguler dan Pembeli Sese kali menunjukkan tumpang tindih yang lebih besar dibandingkan segmen lain, terlihat dari kerapatan titik pada rentang skor kepuasan enam hingga tujuh pada Gambar 4. Kondisi ini mengindikasikan bahwa batas antara dua segmen tersebut bersifat cair, sehingga pelanggan pada zona tumpang tindih ini merupakan kandidat paling potensial untuk program upselling, karena intervensi kepuasan yang kecil saja berpeluang menggeser mereka dari Pembeli Sese kali menuju Pembeli Reguler. Wawasan semacam ini tidak dapat diperoleh dari tabel ringkasan rata-rata segmen saja dan menegaskan pentingnya visualisasi pada level individu, bukan hanya agregat.

4.2 Model Prediksi Churn dan Faktor Penentu

Regresi logistik menunjukkan bahwa skor kepuasan dan tingkat keterlibatan pelanggan merupakan dua prediktor churn paling kuat (p kurang dari 0,01), dengan akurasi model mencapai 82 persen pada data pengujian. Gambar 5 menegaskan pola ini melalui matriks korelasi, di mana skor kepuasan berkorelasi negatif kuat dengan churn dan positif kuat dengan

pengeluaran bulanan, sementara frekuensi transaksi memberikan kontribusi yang lebih moderat.



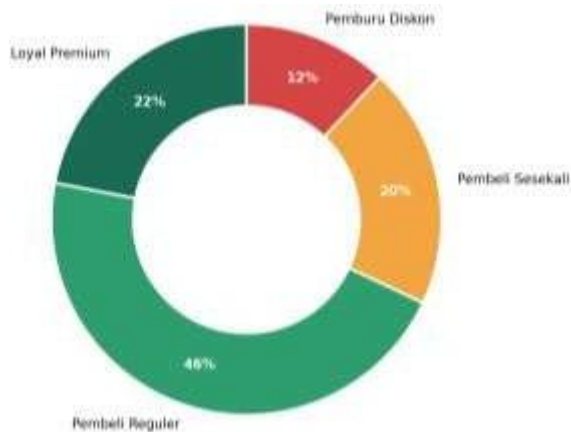
Gambar 5. Korelasi Antar Variabel Perilaku Pelanggan

Temuan ini sejalan dengan argumen Wamba dkk. (2017) bahwa kapabilitas analitik berbasis data meningkatkan kinerja perusahaan melalui kapabilitas dinamis organisasi, tetapi penelitian ini memperluas argumen tersebut pada level operasional yang lebih spesifik, yaitu menunjukkan variabel mana yang paling layak dijadikan pemicu intervensi. Skor kepuasan, bukan sekadar frekuensi transaksi yang selama ini lebih mudah diukur namun kurang prediktif terhadap churn, terbukti menjadi pengungkit utama dalam model ini. Perlu dicatat bahwa akurasi 82 persen tidak serta-merta berarti model ini siap diterapkan tanpa pengawasan lanjutan. Model regresi logistik cenderung kurang sensitif terhadap perubahan perilaku yang terjadi tiba-tiba, misalnya akibat insiden layanan pelanggan atau perubahan kebijakan pengiriman, karena model dilatih pada pola historis. Keterbatasan ini mempertegas argumen pada bagian tinjauan pustaka bahwa kerangka BDA-CRM idealnya bersifat siklikal dengan umpan balik berkelanjutan, sebagaimana digambarkan pada Gambar 2, bukan model statis yang dilatih sekali dan digunakan tanpa pembaruan.

4.3 Komposisi Basis Pelanggan dan Implikasi Anggaran

Dari sisi komposisi, Pembeli Reguler mendominasi basis pelanggan dengan 46 persen,

diikuti Loyal Premium 22 persen dan Pembeli Sesekali 20 persen, sementara Pemburu Diskon menyumbang porsi terkecil, 12 persen, namun memiliki risiko churn tertinggi, sebagaimana digambarkan pada Gambar 6.



Gambar 6. Proporsi Pengguna per Segmen Pelanggan

Komposisi ini mengungkap ketegangan yang jarang dibahas eksplisit dalam literatur BDA-CRM, yakni segmen dengan populasi terbesar (Pembeli Reguler) bukanlah segmen dengan risiko churn tertinggi maupun segmen dengan nilai transaksi tertinggi. Alokasi anggaran promosi yang proporsional terhadap ukuran populasi, sebagaimana lazim dilakukan dalam praktik pemasaran konvensional, berisiko salah sasaran karena mengabaikan bobot risiko dan nilai finansial tiap segmen.

Merujuk pada temuan Wamba dkk. (2017) bahwa kapabilitas analitik meningkatkan kemampuan perusahaan merespons pasar, temuan penelitian ini menunjukkan secara lebih spesifik bahwa respons tersebut seharusnya diarahkan pada pembobotan ulang anggaran retensi, dari alokasi proporsional terhadap populasi menjadi alokasi proporsional terhadap perkalian antara risiko churn dan nilai pelanggan.

4.4 Kontribusi dan Insight Utama

Tiga kontribusi utama penelitian ini dapat dirumuskan. Pertama, penelitian ini menunjukkan secara kuantitatif bahwa kepuasan pelanggan, bukan frekuensi transaksi, adalah pengungkit utama retensi pada konteks e-commerce Indonesia, sebuah temuan yang lebih tajam

dibandingkan simpulan umum pada tinjauan literatur BDA-CRM sebelumnya (Del Vecchio dkk., 2022; Akter & Wamba, 2016). Kedua, alur simulasi dan pemodelan yang dibangun sepenuhnya dengan Python bersifat terbuka dan dapat direproduksi, mengoreksi keterbatasan replikabilitas yang melekat pada studi berbasis tinjauan literatur (Khan dkk., 2024) maupun studi empiris yang berhenti pada tahap segmentasi (Anitha & Patil, 2022; Prakasa & Subardono, 2022). Ketiga, penelitian ini menempatkan kepatuhan terhadap UU PDP 2022 sebagai bagian yang menyatu dengan kerangka pemodelan, bukan sekadar catatan risiko di bagian penutup, sehingga rekomendasi yang dihasilkan lebih siap diterapkan pada konteks regulasi Indonesia dibandingkan rujukan literatur global yang bersifat umum (Roche & Jamal, 2021).

4.5 Batasan Tata Kelola Data dalam Penerapan Praktis

Penerapan keempat langkah di atas tidak dapat dilepaskan dari batasan yang diatur UU PDP 2022, khususnya prinsip pembatasan tujuan pemrosesan data dan kewajiban memperoleh persetujuan eksplisit dari pemilik data untuk aktivitas pemasaran yang bersifat personal. Sistem peringatan dini berbasis skor churn, misalnya, berpotensi bersinggungan dengan hak subjek data untuk tidak diprofilkan secara otomatis apabila tidak disertai mekanisme transparansi yang memadai. Roche dan Jamal (2021) menyebut tantangan semacam ini sebagai persoalan etika yang bersifat umum dalam big data, tetapi pada konteks Tokopedia yang beroperasi di bawah yurisdiksi hukum Indonesia, kepatuhan ini bersifat mengikat secara hukum, bukan sekadar praktik etis yang bersifat sukarela. Implikasinya, tim data science tidak dapat merancang model churn secara terpisah dari tim legal dan kepatuhan, karena kedua fungsi tersebut kini menjadi satu kesatuan proses dalam siklus BDA-CRM.

Selain aspek regulasi, kesenjangan kompetensi analitik pada sumber daya manusia tetap menjadi kendala yang nyata. Model yang akurat secara statistik tidak akan memberi dampak apabila tim pemasaran tidak mampu menerjemahkan koefisien regresi atau titik pusat kluster menjadi keputusan anggaran yang konkret.

Ahmed dkk. (2020) menunjukkan bahwa

infrastruktur pemrosesan seperti Hadoop dan Spark sudah cukup matang untuk menangani skala data e-commerce besar, sehingga hambatan yang tersisa memang lebih banyak terletak pada sisi manusia dan proses organisasi dibandingkan pada sisi teknologi itu sendiri.

5. Kesimpulan

Penelitian ini menjawab tiga tujuan yang diajukan di awal. Pertama, simulasi dan pemodelan Big Data Analytics pada CRM Tokopedia terbukti mampu mengubah pengambilan keputusan dari pendekatan reaktif menjadi prediktif, dengan model regresi logistik mencapai akurasi 82 persen dalam memodelkan churn. Kedua, pemodelan klasterisasi K-Means berhasil mengidentifikasi empat segmen pelanggan dengan karakteristik pengeluaran, kepuasan, dan risiko churn yang berbeda signifikan, memberikan dasar segmentasi yang lebih tajam dibandingkan pendekatan demografis konvensional. Ketiga, hasil pemodelan menunjukkan bahwa skor kepuasan pelanggan adalah prediktor churn paling kuat, sehingga strategi retensi yang efektif perlu diarahkan pada intervensi kepuasan layanan, bukan semata insentif harga.

Implementasi BDA dalam sistem CRM secara praktis bergantung pada tiga pilar utama: tata kelola data yang patuh pada UU PDP No. 27 Tahun 2022, kompetensi analitik sumber daya manusia dalam menerjemahkan luaran model menjadi kebijakan bisnis, serta keselarasan antara strategi retensi dengan aksi eksekusi operasional.

Keterbatasan penelitian ini terletak pada penggunaan data simulasi yang dikalibrasi berdasarkan literatur dan tren sekunder, bukan data transaksi riil dari e-commerce, sehingga generalisasi hasil perlu dilakukan secara hati-hati. Penelitian selanjutnya disarankan memperluas kerangka simulasi dan pemodelan ini melalui kerja sama data dengan platform e-commerce untuk validasi pada data riil, serta membandingkan pola retensi lintas platform di Asia Tenggara guna menilai konsistensi temuan mengenai dominasi variabel kepuasan terhadap churn.

6. Daftar Pustaka

- [1] Ahmed, N., Barczak, A. L. C., Susnjak, T., & Rashid, M. A. (2020). A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench. *Journal of Big Data*, 7, Article 110. <https://doi.org/10.1186/s40537-020-00388-5>
- [2] Akter, S., & Wamba, S. F. (2016). Big data analytics in e-commerce: A systematic review and agenda for future research. *Electronic Markets*, 26(2), 173-194. <https://doi.org/10.1007/s12525-016-0219-0>
- [3] Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5), 1785-1792. <https://doi.org/10.1016/j.jksuci.2019.12.011>
- [4] Bank Indonesia. (2025, Agustus 26). *Transaksi e-commerce Indonesia tembus Rp487 triliun pada 2024* [Data diolah]. Media Kaltim. <https://mediakaltim.com/transaksi-e-commerce-indonesia-tembus-rp487-triliun-pada-2024/>
- [5] Del Vecchio, P., Mele, G., Siachou, E., & Schito, G. (2022). A structured literature review on Big Data for customer relationship management (CRM): Toward a future agenda in international marketing. *International Marketing Review*, 39(5), 1069-1092. <https://doi.org/10.1108/IMR-01-2021-0036>
- [6] Google, Temasek, & Bain & Company. (2025, November 13). *e-Conomy SEA 2025* [Data diolah]. Kompas.com. <https://tekno.kompas.com/read/2025/11/13/19150087/nilai-bisnis-e-commerce-indonesia-tahun-2025-tembus-rp-1187-triliun>
- [7] Khan, M. A. R., Akter, M. S., & Islam, R. (2024). Big data analytics and predictive analysis in enhancing customer relationship management (CRM): A systematic review of techniques and tools. *Journal of Machine Learning, Data Engineering and Data Science*, 1(1), 83-99. <https://doi.org/10.70008/jmldeds.v1i01.44>

- [8] Momentum Works. (2024). *Shopee, TikTok Shop, and Blibli market share in Indonesia 2024* [Data diolah]. Databoks. <https://databoks.katadata.co.id/en/technology--telecommunications/statistics/68629ac92be0b/>
- [9] Republik Indonesia. (2022). *Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi*.
- [10] Roche, J., & Jamal, A. (2021). A systematic literature review of the role of ethics in big data. In H. Jahankhani, A. Jamal, & S. Lawson (Eds.), *Cybersecurity, privacy and freedom protection in the connected world*. Springer. https://doi.org/10.1007/978-3-030-68534-8_20
- [11] Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, 356-365. <https://doi.org/10.1016/j.jbusres.2016.08.009>