



Prediksi *Churn* Pelanggan Telekomunikasi Menggunakan Segmentasi Pelanggan dan *Explainable Artificial Intelligence* Berbasis SHAP

Ika Kemala Sawati Azzahra^{1*}, I Made Artha Agastya²

¹⁻² Program Studi Informatika, Universitas Amikom Yogyakarta, Indonesia

Email: ikakemalasa@students.amikom.ac.id^{1*}, artha.agastya@amikom.ac.id²

*Penulis Korespondensi: ikakemalasa@students.amikom.ac.id

Abstract. *The telecommunications industry faces intense competition that makes customer churn a critical threat to business sustainability. This study proposes an integrated analytical framework that combines churn prediction, customer segmentation, and model interpretability to support data-driven retention strategies. Three machine learning algorithms Logistic Regression, Random Forest, and XGBoost were trained on the IBM Telco dataset (n = 7,043) with class imbalance handled through SMOTE. K-Means clustering was then applied to predicted-churn customers, and SHAP (SHapley Additive exPlanations) was used to interpret the best-performing model both globally and at the segment level. Random Forest achieved the best performance (Accuracy 0.7763, F1-Score 0.6114, ROC-AUC 0.833), and three customer segments were identified: New High-Risk, Loyal High-Value, and Low-Tier Economic. The per-segment SHAP analysis revealed heterogeneous churn drivers across segments, enabling the formulation of more targeted and actionable retention strategies. The integration of churn prediction, segmentation, and SHAP therefore provides a more precise foundation for data-driven retention decision-making in the telecommunications industry.*

Keywords: *Customer Churn; Customer Segmentation; K-Means; Machine Learning; SHapley Additive exPlanations.*

Abstrak. Industri telekomunikasi menghadapi persaingan yang sangat ketat sehingga customer churn menjadi ancaman serius bagi keberlanjutan bisnis. Penelitian ini mengusulkan kerangka analitik terpadu yang menggabungkan prediksi churn, segmentasi pelanggan, dan analisis interpretabilitas model untuk mendukung strategi retensi berbasis data. Tiga algoritma machine learning Logistic Regression, Random Forest, dan XGBoost dilatih pada dataset IBM Telco (n = 7.043) dengan penanganan ketidakseimbangan kelas menggunakan SMOTE. Selanjutnya, K-Means clustering diterapkan pada pelanggan yang diprediksi churn, dan SHAP (SHapley Additive exPlanations) digunakan untuk menginterpretasikan model terbaik baik secara global maupun pada tingkat segmen. Random Forest memberikan performa terbaik (Accuracy 0.7763, F1-Score 0.6114, ROC-AUC 0.833), dan diperoleh tiga segmen pelanggan: Pelanggan Baru Rentan Pergi, Pelanggan Loyal Bernilai Tinggi, serta Pelanggan Hemat. Analisis SHAP per segmen mengungkap heterogenitas faktor pendorong churn antar kelompok pelanggan sehingga strategi retensi dapat dirumuskan secara lebih spesifik dan tepat sasaran. Integrasi prediksi churn, segmentasi, dan SHAP terbukti memberikan dasar pengambilan keputusan retensi yang lebih presisi dan dapat ditindaklanjuti dalam industri telekomunikasi.

Kata kunci: *K-Means; Pembelajaran Mesin; Pengurangan Pelanggan; Penjelasan Aditif SHapley; Segmentasi Pelanggan.*

1. LATAR BELAKANG

Industri telekomunikasi merupakan salah satu sektor dengan tingkat persaingan tertinggi di era digital. Berbagai operator berlomba menawarkan layanan yang lebih baik dengan harga yang lebih kompetitif, sehingga perpindahan pelanggan (customer churn) menjadi ancaman signifikan bagi keberlanjutan bisnis (Lalwani et al., 2022). Biaya untuk mendapatkan pelanggan baru diperkirakan lima hingga sepuluh kali lebih besar dibandingkan biaya mempertahankan pelanggan yang sudah ada (Wu et al., 2021). Oleh karena itu, kemampuan untuk memprediksi dan mencegah churn memiliki nilai strategis yang tinggi.

Prediksi churn menggunakan machine learning telah banyak diteliti dan terbukti mampu mengidentifikasi pelanggan yang berpotensi meninggalkan layanan (Adeniran et al., 2024). Tinjauan sistematis terbaru oleh Onaja Paul. J. (Onoja, 2025) terhadap publikasi pada periode 2020–2024 menyimpulkan bahwa algoritma berbasis *ensemble* dan *deep learning* mendominasi penelitian churn telekomunikasi, dengan SMOTE sebagai teknik penyeimbangan data paling banyak digunakan. Berbagai algoritma telah diuji, mulai dari Logistic Regression, Decision Tree, Random Forest, hingga pendekatan ensemble seperti XGBoost (Sikri et al., 2024). Namun, sebagian besar penelitian hanya berfokus pada peningkatan akurasi prediksi tanpa memberikan penjelasan tentang mengapa seorang pelanggan cenderung churn (Leung et al., 2021).

Di sisi lain, customer segmentation memberikan pemahaman tentang karakteristik kelompok pelanggan, yang berguna untuk personalisasi layanan dan strategi pemasaran (Harani et al., 2020). Meskipun beberapa penelitian telah menggabungkan prediksi churn dengan segmentasi pelanggan (Wu et al., 2021), analisis faktor churn yang spesifik per segmen masih jarang dilakukan. Akibatnya, strategi retensi yang dihasilkan cenderung bersifat umum dan tidak tepat sasaran.

Explainable Artificial Intelligence (XAI) hadir sebagai solusi untuk membuka "kotak hitam" model machine learning, sehingga keputusan model dapat dipahami oleh pengambil kebijakan (Bauer et al., 2023). Salah satu metode XAI yang paling banyak digunakan adalah SHAP (SHapley Additive exPlanation), yang mampu mengukur kontribusi setiap fitur terhadap prediksi model, baik secara global maupun lokal (Esfahani et al., 2025). Chang et al. (Chang et al., 2024) menunjukkan bahwa kombinasi Random Forest dengan SHAP menghasilkan model prediksi churn yang tidak hanya akurat tetapi juga interpretatif.

Namun demikian, penelitian yang mengintegrasikan prediksi churn, segmentasi pelanggan, dan analisis SHAP per segmen masih sangat terbatas. Wu et al. (Wu et al., 2021) menggabungkan churn prediction dengan K-Means clustering, tetapi menggunakan Bayesian Analysis sebagai pendekatan interpretabilitas yang kurang visual dan kurang informatif dibandingkan SHAP. Penelitian ini mengisi kesenjangan tersebut dengan mengusulkan kerangka analitik terpadu yang menggabungkan ketiga komponen tersebut.

Kontribusi utama penelitian ini adalah: (a) Mengintegrasikan prediksi churn, segmentasi pelanggan (K-Means), dan analisis SHAP dalam satu pipeline yang kohesif, (b) Menyajikan analisis SHAP per segmen, bukan hanya secara global, sehingga faktor penyebab churn dapat dipahami untuk setiap kelompok pelanggan, (c) Merumuskan rekomendasi strategi retensi yang spesifik untuk masing-masing segmen pelanggan berdasarkan temuan SHAP.

2. KAJIAN TEORITIS

Prediksi Churn Pelanggan

Prediksi churn pelanggan merupakan masalah klasifikasi biner di mana tujuannya adalah mengidentifikasi pelanggan yang berpotensi berhenti menggunakan layanan. Berbagai penelitian telah mengevaluasi kinerja algoritma machine learning untuk tugas ini. Chang et al. (Chang et al., 2024) menggunakan ensemble learning seperti Decision Tree, Random Forest, dan Logistic Regression pada dataset telekomunikasi publik, dengan Random Forest mencapai akurasi terbaik sebesar 91.66%. Sikri et al. (Sikri et al., 2024) mengusulkan teknik Ratio-Based Data Balancing yang dikombinasikan dengan XGBoost, menghasilkan akurasi 89.6% dengan F1-Score 68.8% pada rasio 75:25. Lalwani et al. (Lalwani et al., 2022) membandingkan delapan algoritma dengan AdaBoost dan XGBoost sebagai performa terbaik F1-Score 81%.

Penelitian-penelitian tersebut menunjukkan bahwa algoritma ensemble seperti Random Forest dan XGBoost secara konsisten menghasilkan performa yang lebih baik dibandingkan algoritma tunggal. Selain itu, penanganan ketidakseimbangan kelas merupakan aspek penting karena proporsi pelanggan yang churn umumnya jauh lebih kecil dibandingkan yang tidak churn.

Penanganan Ketidakseimbangan Data

Dataset churn pelanggan umumnya tidak seimbang, di mana kelas minoritas (churn) jauh lebih sedikit dibandingkan kelas mayoritas (tidak churn). Kondisi ini dapat menyebabkan model bias terhadap kelas mayoritas sehingga menghasilkan recall yang rendah untuk kelas churn (Sikri et al., 2024). SMOTE (*Synthetic Minority Over-sampling Technique*) merupakan teknik *oversampling* paling umum untuk mengatasi masalah ini dengan membuat sampel sintetis pada kelas minoritas berdasarkan interpolasi tetangga terdekat (Mahmud et al., 2024; Wu et al., 2021). Variasi modern seperti Borderline-SMOTE dan ADASYN juga banyak digunakan, namun SMOTE klasik tetap menjadi *baseline* yang relevan karena kestabilan dan kesederhanaannya. Imani et al. (Imani et al., 2024) menemukan bahwa kombinasi SMOTE dengan Random Forest dan Gradient Boosting menghasilkan F1-Score lebih tinggi dibandingkan ADASYN pada dataset churn telekomunikasi. Salehi (Salehi & Khedmati, 2024) selanjutnya memperkenalkan algoritma *Cluster-based SMOTE Both-sampling* (CSBBoost) yang menggabungkan *clustering* dengan SMOTE untuk meningkatkan kualitas sampel sintetis pada kelas minoritas.

Segmentasi Pelanggan

Segmentasi pelanggan bertujuan mengelompokkan pelanggan berdasarkan kesamaan karakteristik untuk mendukung strategi pemasaran yang lebih tepat sasaran. K-Means merupakan algoritma clustering yang paling banyak digunakan karena sederhana, efisien, dan mudah diinterpretasikan (Harani et al., 2020). Wu et al. (Wu et al., 2021) menggunakan K-Means untuk mengelompokkan pelanggan churn menjadi tiga cluster berdasarkan fitur-fitur yang dipilih melalui analisis Bayesian, dan menemukan bahwa karakteristik churn berbeda secara signifikan antar cluster. Penentuan jumlah cluster optimal lazimnya menggunakan kombinasi Elbow Method, Silhouette Score.

Explainable Artificial Intelligence (XAI)

Model machine learning yang kompleks seperti Random Forest dan XGBoost sering dianggap sebagai black box karena mekanisme internalnya sulit dipahami (Bauer et al., 2023). Explainable AI (XAI) hadir untuk meningkatkan transparansi dan interpretabilitas model. Dua metode XAI yang paling banyak digunakan adalah LIME dan SHAP (Chang et al., 2024). SHAP (SHapley Additive exPlanation) mengukur kontribusi setiap fitur terhadap prediksi model berdasarkan teori permainan Shapley (Esfahani et al., 2025). Nilai SHAP untuk fitur i didefinisikan sebagai:

$$\sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad \dots\dots(i)$$

Di mana F adalah himpunan seluruh fitur, S adalah subset fitur yang tidak memuat fitur i , dan $f(S)$ adalah prediksi model menggunakan subset S . Keunggulan SHAP terletak pada kemampuannya memberikan penjelasan baik secara global (keseluruhan dataset) maupun lokal (per sampel individu) dengan jaminan teoretis berupa *local accuracy*, *missingness*, dan *consistency*. Chang et al. (Chang et al., 2024) menggunakan LIME dan SHAP untuk menginterpretasikan Random Forest, dan menemukan bahwa fitur *Contract*, *Number of Referrals*, *Tenure in Months*, *Monthly Charge*, dan *Online Security* adalah faktor dominan penyebab churn. Noviandy et al. (Noviandy et al., 2024) mengusulkan kerangka *tabular machine learning* dengan Gradient Boosting Machine pada dataset IBM Telco yang menghasilkan akurasi 80.70% disertai visualisasi SHAP untuk menjelaskan prediksi pada level global maupun lokal. Özkurt (Özkurt, 2025) melakukan analisis komparatif terhadap dua teknik XAI, yaitu SHAP dan *Interpreted Machine Learning Partial Dependence*, pada dataset Telco Customer Churn ($n=7.043$), dan menyimpulkan bahwa SHAP secara konsisten memberikan penjelasan yang lebih informatif untuk pengambilan keputusan retensi.

Posisi Penelitian

Tabel 1 merangkum perbandingan penelitian terkait dengan penelitian ini.

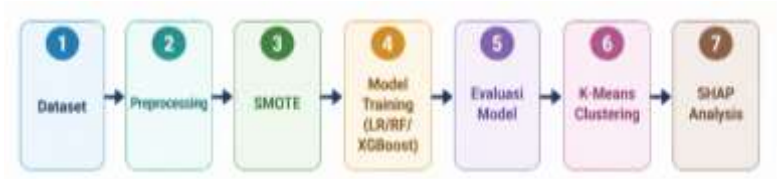
Tabel 1. Komparasi dengan Studi Terdahulu.

Studi	Algoritma	Penanganan Imbalance	Segmentasi	Interpretabilitas	Analisis per Segmen
Chang et al (Chang et al., 2024)	Logistic Regression, KNN, Naïve Bayes, Decision Tree, Random Forest	Tidak ada	Tidak ada	LIME + SHAP (lokal dan global)	Analisis 2 sampel individual via LIME/SHAP
Wu et al (Wu et al., 2021)	LR, Decision Tree, Random Forest, Naïve Bayes, AdaBoost, MLP	SMOTE	K-Means (3 cluster)	Bayesian Analysis	Sebagian
Penelitian ini	Logistic Regression, Random Forest, XGBoost	SMOTE	K-Means (multi-metric)	SHAP Global + Per Segmen	Ya

Posisi diferensial penelitian ini terletak pada kombinasi SHAP per segmen dengan validasi statistik antar cluster, yang belum ditemukan pada studi-studi sebelumnya.

3. METODE PENELITIAN

Gambar 1 mengilustrasikan pipeline metodologi penelitian ini yang terdiri dari tujuh tahap utama: pemrosesan data, penyeimbangan kelas, *training* model, evaluasi, segmentasi, analisis SHAP global, dan analisis SHAP per segmen.



Gambar 1. Diagram Alur Metodologi Penelitian.

Dataset

Penelitian ini menggunakan dataset IBM Telco Customer Churn yang tersedia secara publik melalui platform Kaggle (IBM, 2018). Dataset ini terdiri dari 7.043 baris data pelanggan dengan 21 atribut, termasuk satu variabel target (Churn: Yes/No). Fitur-fitur dalam dataset meliputi informasi demografis pelanggan (gender, SeniorCitizen, Partner, Dependents), informasi layanan yang digunakan (PhoneService, InternetService, OnlineSecurity, TechSupport, dll.), serta informasi kontrak dan pembayaran (Contract, PaymentMethod, MonthlyCharges, TotalCharges). Distribusi kelas dalam dataset menunjukkan ketidakseimbangan, dengan 1869 pelanggan yang churn (26.5%) dan 5174 yang tidak churn (73.5%).

Preprocessing

Tahap preprocessing meliputi: (a) Penghapusan atribut customerID yang tidak relevan untuk prediksi. (b) Konversi atribut TotalCharges ke tipe data numerik dan penanganan nilai kosong dengan imputasi nilai median (c) Konversi label target Churn dari Yes/No menjadi 1/0 (d) Label Encoding untuk seluruh variabel kategorik.

Penanganan Ketidakseimbangan Data (SMOTE)

SMOTE diterapkan pada data training untuk menyeimbangkan distribusi kelas. SMOTE bekerja dengan membuat sampel sintetis pada kelas minoritas berdasarkan interpolasi di antara sampel-sampel yang berdekatan dalam ruang fitur. Proporsi pembagian data adalah 75% training dan 25% testing dengan stratified split untuk mempertahankan proporsi kelas pada kedua set.

Model Prediksi Churn

Tiga model machine learning digunakan dalam penelitian ini: (a) Logistic Regression: Digunakan sebagai baseline. Model linear ini menggunakan fungsi logit untuk menghasilkan probabilitas churn. (b) Random Forest: Algoritma ensemble berbasis bagging yang membangun sejumlah pohon keputusan secara paralel dan mengambil voting mayoritas sebagai prediksi akhir. Parameter yang digunakan: `n_estimators=200`, `max_depth=10`. (c) XGBoost (Extreme Gradient Boosting): Algoritma ensemble berbasis boosting yang membangun pohon keputusan secara sekuensial, di mana setiap pohon baru memperbaiki kesalahan pohon sebelumnya. Parameter utama: `n_estimators=200`, `max_depth=6`, `learning_rate=0.05`.

Metrik Evaluasi

Performa model dievaluasi menggunakan lima metrik: Accuracy, Precision, Recall, F1-Score, dan ROC-AUC. Untuk masalah churn prediction dengan data tidak seimbang, Recall dan F1-Score menjadi metrik yang lebih penting dibandingkan Accuracy, karena operator lebih mengutamakan kemampuan mengidentifikasi pelanggan yang benar-benar akan churn (meminimalkan False Negative).

Segmentasi Pelanggan (K-Means)

Segmentasi dilakukan pada data pelanggan yang diprediksi akan churn. Fitur yang digunakan untuk clustering adalah: tenure, MonthlyCharges, TotalCharges, Contract, InternetService, PaymentMethod, OnlineSecurity, TechSupport, dan OnlineBackup. Data dinormalisasi menggunakan StandardScaler sebelum clustering.

Jumlah cluster optimal (K) ditentukan menggunakan dua metode: (1) Elbow Method yang mengamati penurunan inertia (Sum of Squared Errors), dan (2) Silhouette Score yang

mengukur seberapa mirip setiap objek dengan cluster-nya sendiri dibandingkan dengan cluster lain.

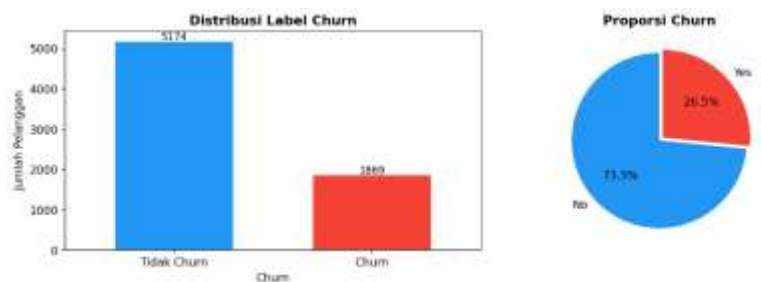
Analisis SHAP (Global dan Per Segmen)

SHAP TreeExplainer digunakan untuk menghasilkan nilai SHAP bagi model terbaik. Analisis dilakukan pada dua level: (a) Global: Menghitung mean absolute SHAP value untuk seluruh test set, menghasilkan pemahaman tentang fitur-fitur yang paling berpengaruh secara keseluruhan. (b) Per Segmen: Menghitung nilai SHAP secara terpisah untuk setiap cluster K-Means, sehingga faktor dominan penyebab churn dapat diidentifikasi untuk masing-masing kelompok pelanggan. Ini merupakan kontribusi utama penelitian ini.

4. HASIL DAN PEMBAHASAN

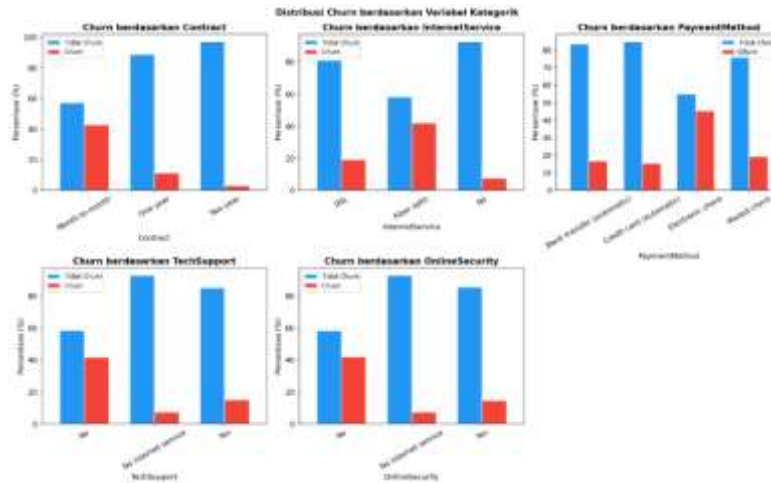
Exploratory Data Analysis (EDA)

Dataset IBM Telco terdiri dari 7.043 pelanggan dengan 26.5% di antaranya (1.869 pelanggan) merupakan pelanggan churn. Distribusi ini menunjukkan ketidakseimbangan kelas yang perlu ditangani sebelum pemodelan.



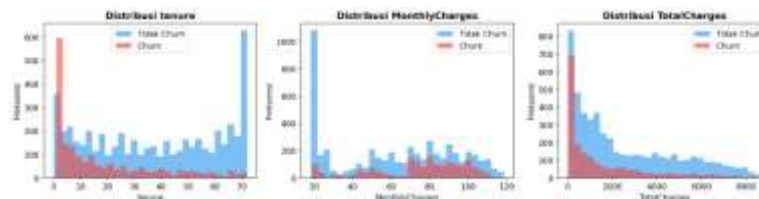
Gambar 2. Distribusi label churn (bar chart dan pie chart).

Analisis EDA terhadap variabel kategorik menunjukkan beberapa temuan penting, pelanggan dengan kontrak bulan-ke-bulan (Month-to-Month) memiliki tingkat churn tertinggi, pelanggan tanpa layanan Online Security dan Tech Support cenderung lebih sering churn, serta pelanggan yang menggunakan Fiber Optic sebagai layanan internet memiliki tingkat churn lebih tinggi dibandingkan DSL.



Gambar 3. Distribusi churn berdasarkan variabel kategorik.

Untuk variabel numerik, pelanggan churn cenderung memiliki tenure yang lebih pendek, Monthly Charges yang lebih tinggi, dan Total Charges yang lebih rendah.



Gambar 4. Distribusi variabel numerik (tenure, MonthlyCharges, TotalCharges).

Preprocessing dan SMOTE

Setelah preprocessing, dataset bersih terdiri dari 7043 baris dengan 20 fitur prediksi. Pembagian data menghasilkan 5282 sampel training dan 1761 sampel testing. Sebelum SMOTE, distribusi kelas pada training set adalah: (a) Kelas 0 (Tidak Churn): 3880 sampel. (b) Kelas 1 (Churn): 1402 sampel. Setelah SMOTE, kedua kelas menjadi seimbang dengan masing-masing 3880 sampel, sehingga model dapat belajar karakteristik churn secara lebih efektif.

Perbandingan Performa Model

Tabel 2 menyajikan perbandingan performa ketiga model yang diuji.

Tabel 2. Perbandingan Performa Model Prediksi Churn.

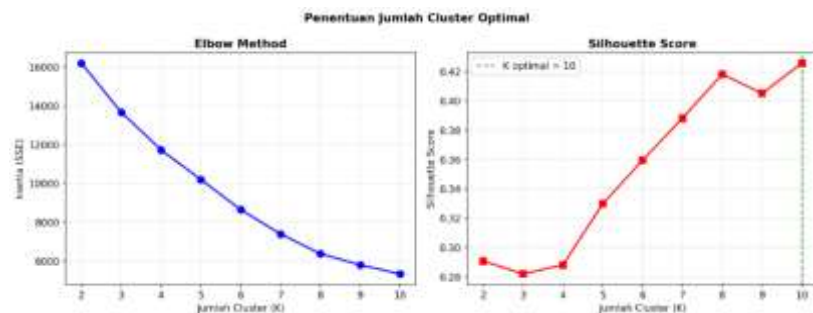
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.7581	0.5326	0.7173	0.6113	0.827
Random Forest	0.7763	0.5667	0.6638	0.6114	0.833
XGBoost	0.7746	0.5711	0.6017	0.5860	0.833

Berdasarkan hasil evaluasi, Random Forest memberikan performa terbaik dengan F1-Score 0.6114 dan ROC-AUC 0.833. Hal ini konsisten dengan temuan Chang et al. (Chang et al., 2024) yang menemukan bahwa ensemble learning memberikan akurasi lebih tinggi dibandingkan algoritma tunggal pada dataset IBM Telco.

Penerapan SMOTE secara konsisten meningkatkan nilai Recall dibandingkan tanpa SMOTE, yang berarti model lebih mampu mengidentifikasi pelanggan yang benar-benar akan churn. Meskipun SMOTE sedikit menurunkan Accuracy, peningkatan Recall sangat penting dalam konteks churn management di mana melewatkan pelanggan yang akan churn (False Negative) jauh lebih merugikan dibandingkan kesalahan mengklasifikasikan pelanggan yang tidak akan churn (Wu et al., 2021).

Hasil Segmentasi Pelanggan (K-Means)

Penentuan jumlah cluster optimal dilakukan melalui pendekatan multi-kriteria yang menggabungkan metrik kuantitatif (Elbow Method, Silhouette Score) dengan pertimbangan interpretabilitas bisnis. Analisis dilakukan pada subset pelanggan yang diprediksi churn ($n = 2.261$).



Gambar 5. Penentuan Jumlah Cluster Optimal melalui Elbow Method dan Silhouette Score.

Hasil Elbow Method menunjukkan penurunan inertia (Sum of Squared Errors) yang tajam pada rentang $K=2$ hingga $K=4$, kemudian melandai secara bertahap. Titik infleksi (elbow point) teridentifikasi pada $K=3$, yang mengindikasikan bahwa penambahan cluster di luar titik tersebut memberikan reduksi varians intra-cluster yang semakin marginal. Sebaliknya, Silhouette Score memuncak pada $K=10$ dengan nilai 0.42, namun tren tersebut menyiratkan pemecahan data ke dalam segmen yang sangat kecil dan secara substantif sulit ditafsirkan untuk keperluan strategi retensi.

Pemilihan $K=3$ didasarkan pada tiga argumentasi yang saling melengkapi. Pertama, prinsip parsimony dalam segmentasi pelanggan menuntut keseimbangan antara kekuatan struktur statistik dan keterterapan operasional. Cluster dengan jumlah anggota terlalu kecil (sebagaimana akan dihasilkan pada $K=8-10$) tidak ekonomis bagi operator karena biaya perancangan kampanye retensi spesifik per segmen melampaui potensi manfaatnya. Kedua, konvergensi antara hasil Elbow Method ($K=3$) dan literatur sebelumnya memperkuat validitas pilihan ini. Wu et al. (Wu et al., 2021) juga menetapkan tiga cluster optimal pada konteks segmentasi pelanggan churn telekomunikasi, sehingga $K=3$ memiliki dukungan teoretis maupun empiris dari penelitian terdahulu pada domain yang setara. Ketiga, pertimbangan

interpretabilitas bisnis: tiga cluster memungkinkan pemetaan yang jelas terhadap tipologi klasik pelanggan telekomunikasi, yaitu new high-risk, loyal high-value, dan low-tier economic kategori yang lazim digunakan dalam praktik Customer Relationship Management.

Dengan $K=3$, diperoleh Silhouette Score sebesar 0.2819 yang tergolong moderate weak structure menurut kriteria Kaufman & Rousseeuw (rentang 0.25–0.5). Nilai ini diakui sebagai keterbatasan, namun dapat dipertanggungjawabkan mengingat dataset memuat fitur kategorikal yang mendominasi (Contract, InternetService, OnlineSecurity, TechSupport, OnlineBackup, PaymentMethod) sehingga ruang fitur cenderung diskrit dan tidak ideal untuk algoritma berbasis jarak Euclidean. Tabel 3 merangkum karakteristik rata-rata masing-masing cluster setelah pemilihan $K=3$ ditetapkan.

Tabel 3. Karakteristik Rata-rata Tiap Cluster Pelanggan Churn.

Fitur	Cluster 0	Cluster 1	Cluster 2
N (Pelanggan)	1325	462	474
Tenure (Bulan)	8.67	43.47	6.27
MonthlyCharges	79.03	93.79	45.26
TotalCharges	728.13	4083.43	285.57
Contract (Encode)	0.00	0.09	0.02
InternetService	1.07	0.97	0.00
PaymentMethod	1.88	1.43	2.00
OnlineSecurity	0.20	0.17	0.19
TechSupport	0.23	0.33	0.19
OnlineBackup	0.41	1.00	0.23

Berdasarkan karakteristik tersebut, ketiga cluster dapat diinterpretasikan sebagai berikut: (a) Cluster 0: "Pelanggan Baru Rentan Pergi" Kelompok ini ditandai dengan Tenure pendek, MonthlyCharges tinggi, dan penggunaan FiberOptic dengan tagihan bulanan besar, namun belum loyal dan minim layanan tambahan. Pelanggan dalam cluster ini merupakan pelanggan baru bernilai tinggi tapi rawan berhenti. (b) Cluster 1: "Pelanggan Loyal Bernilai Tinggi" Kelompok ini ditandai dengan Tenure Panjang, TotalCharges tertinggi, MonthlyCharges Tertinggi, dan kontrak cendrung jangka panjang, menggunakan layanan tambahan paling lengkap. Pelanggan ini merupakan tulang punggung pendapatan perusahaan. (c) Cluster 2: "Pelanggan Hemat" Kelompok ini ditandai dengan Tenure sangat pendek, MonthlyCharges rendah, TotalCharges paling rendah, kemungkinan hanya layanan dasar. Pelanggan ini merupakan pelanggan ekonomis dengan kontribusi pendapatan kecil.

Analisis SHAP Global

Analisis SHAP dilakukan menggunakan Random Forest sebagai model terbaik. Tabel 4 menunjukkan 10 fitur teratas berdasarkan mean absolute SHAP value secara global.

Tabel 4. Top 10 Fitur SHAP Global.

Peringkat	Feature	Mean SHAP
1	Contract	0.120800
2	OnlineSecurity	0.068543
3	MonthlyCharges	0.057201
4	TechSupport	0.052608
5	Tenure	0.049462
6	InternetService	0.037065
7	OnlineBackup	0.031829
8	Dependents	0.029140
9	TotalCharges	0.023290
10	Partner	0.019296

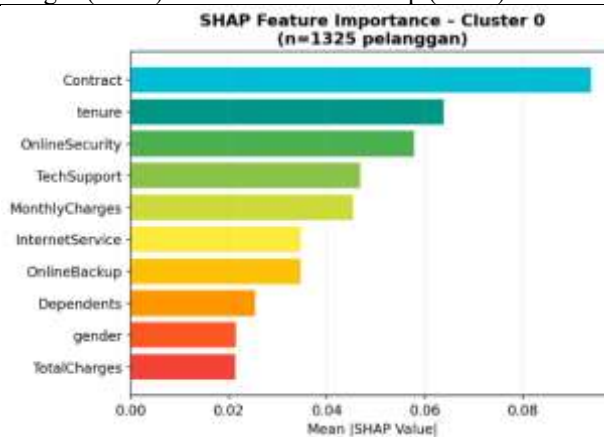
Secara global, fitur Contract, OnlineSecurity, dan MonthlyCharges merupakan tiga faktor yang paling berpengaruh terhadap prediksi churn. Temuan ini sejalan dengan penelitian Chang et al.(Chang et al., 2024) yang menemukan bahwa Contract, Tenure in Months, dan Monthly Charge merupakan faktor dominan pada dataset IBM Telco.

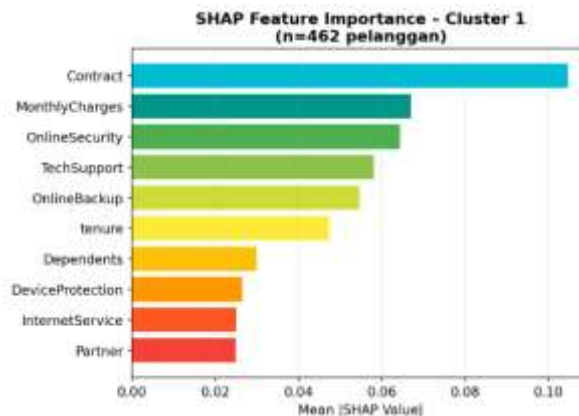
Analisis SHAP Per Segmen

Bagian ini merupakan kontribusi utama penelitian yang membedakannya dari penelitian sebelumnya. Tabel 5 merangkum lima faktor churn dominan untuk masing-masing cluster.

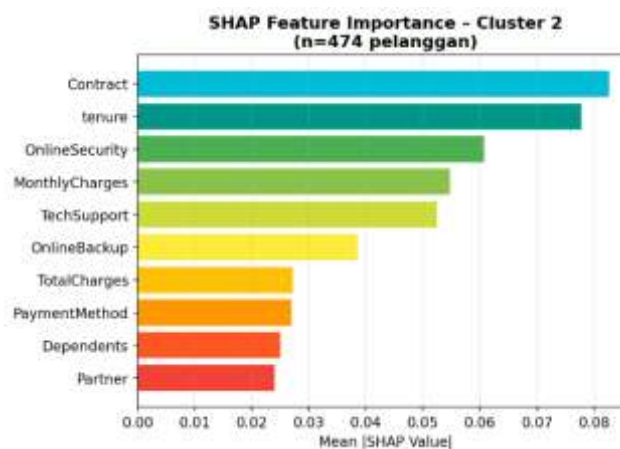
Tabel 5. Top 5 Faktor Churn per Segmen Berdasarkan Analisis SHAP.

Peringkat	Cluster 0 (Mean SHAP)	Cluster 1 (Mean SHAP)	Cluster 2 (Mean SHAP)
1	Contract (0.094)	Contract (0.105)	Contract (0.082)
2	Tenure (0.0653)	MonthlyCharges (0.068)	Tenure (0.078)
3	OnlineSecurity (0.058)	OnlineSecurity (0.062)	OnlineSecurity (0.061)
4	TechSupport (0.049)	TechSupport (0.057)	MonthlyCharges (0.055)
5	MonthlyCharges (0.047)	OnlineBackup (0.053)	TechSupport (0.053)

**Gambar 6.** SHAP Bar Plot untuk Cluster 0.



Gambar 7. SHAP Bar Plot untuk Cluster 1.



Gambar 8. SHAP Bar Plot untuk Cluster 2.

Hasil analisis SHAP per segmen mengungkap perbedaan signifikan pada faktor-faktor pendorong churn antar cluster. Temuan ini mengonfirmasi hipotesis bahwa determinan churn bersifat heterogen pada populasi pelanggan telekomunikasi, sehingga interpretasi global tidak memadai untuk merumuskan kebijakan retensi yang presisi.

Pada Cluster 0 "Pelanggan Baru Rentan Pergi", $n=1325$, fitur Contract memperoleh nilai mean SHAP tertinggi (0.094) diikuti tenure 0.065 dan OnlineSecurity 0.058. Dominasi Contract dan tenure mengindikasikan bahwa keputusan churn pada segmen ini didorong oleh kerentanan struktural kontrak bulan-ke-bulan yang dipadukan dengan masa berlangganan yang masih singkat. Pelanggan belum memiliki sunk cost emosional maupun finansial yang cukup besar, sehingga hambatan untuk berpindah operator menjadi rendah. Tarif bulanan yang relatif tinggi pada segmen ini memperbesar sensitivitas harga, terutama ketika ekspektasi nilai layanan belum terpenuhi pada periode awal berlangganan.

Pada Cluster 1 "Pelanggan Loyal Bernilai Tinggi", $n=462$, fitur Contract tetap mendominasi 0.105, namun urutan kedua ditempati oleh MonthlyCharges 0.068 dan OnlineSecurity 0.062, sementara tenure justru bergeser ke peringkat keenam. Pergeseran ini

menunjukkan bahwa loyalitas yang terbangun melalui masa berlangganan panjang telah menurunkan kontribusi tenure terhadap probabilitas churn, sedangkan beban tagihan bulanan tertinggi di antara seluruh cluster menjadi titik tekan utama. Risiko churn pada segmen ini berakar pada persepsi value-for-money: ketika kenaikan tarif tidak diimbangi peningkatan kualitas layanan keamanan dan pencadangan data, pelanggan bernilai tinggi cenderung mengevaluasi ulang keputusan berlangganannya.

Pada Cluster 2 "Pelanggan Hemat", $n=474$, fitur Contract 0.082, tenure 0.078, dan OnlineSecurity 0.061 menjadi tiga kontributor utama. Pola SHAP pada segmen ini paralel dengan Cluster 0, tetapi dengan magnitudo MonthlyCharges yang lebih rendah karena pelanggan hanya berlangganan layanan dasar tanpa internet. Determinan churn pada kelompok ini lebih banyak dipengaruhi oleh ketiadaan pengikat kontraktual dan minimnya layanan bernilai tambah, bukan oleh tekanan harga.

Rekomendasi Strategi Retensi Per Segmen

Berdasarkan karakteristik cluster pada Tabel 3 serta hasil analisis SHAP per segmen pada Tabel 5, dirumuskan strategi retensi yang disesuaikan dengan profil risiko dan determinan churn masing-masing kelompok pelanggan.

Cluster 0: Pelanggan Baru Rentan Pergi ($n = 1.325$). Segmen ini dicirikan oleh *tenure* pendek (rata-rata 8.67 bulan), MonthlyCharges tinggi (79.03), pengguna Fiber Optic, kontrak *month-to-month*, serta layanan tambahan minim. Faktor churn dominan berturut-turut adalah Contract, *tenure*, OnlineSecurity, TechSupport, dan MonthlyCharges. Strategi retensi yang direkomendasikan meliputi: (1) konversi kontrak bulanan menjadi kontrak satu atau dua tahun melalui diskon progresif 10–20%, (2) *bundling* Fiber Optic dengan OnlineSecurity dan TechSupport secara gratis pada enam bulan pertama, (3) program *onboarding* terstruktur untuk membangun persepsi nilai pada masa kritis 0–6 bulan, serta (4) *early-warning system* berbasis skor probabilitas churn dengan intervensi proaktif *customer care*.

Cluster 1 — Pelanggan Loyal Bernilai Tinggi ($n = 462$). Segmen ini ditandai dengan *tenure* panjang (rata-rata 43.47 bulan), TotalCharges tertinggi (4.083,43), MonthlyCharges tertinggi (93.79), kontrak jangka panjang, dan layanan tambahan paling lengkap. Faktor churn dominan adalah Contract, MonthlyCharges, OnlineSecurity, TechSupport, dan OnlineBackup. Strategi retensi yang direkomendasikan mencakup: (1) penerapan *loyalty tier* dengan rebate tarif progresif berdasarkan masa berlangganan, (2) perpanjangan kontrak otomatis disertai *price-lock guarantee* selama 12–24 bulan, (3) peningkatan layanan premium (*priority support*, *enhanced security*, *expanded backup quota*) tanpa biaya tambahan, serta (4) penugasan *account manager* khusus bagi pelanggan *high-value* sebagai penjaga relasi jangka panjang.

Cluster 2 — Pelanggan Hemat ($n = 474$). Segmen ini memiliki *tenure* sangat pendek (rata-rata 6.27 bulan), *MonthlyCharges* rendah (45.26), *TotalCharges* paling rendah (285.57), tanpa layanan internet, dan hanya menggunakan layanan dasar. Faktor churn dominan adalah *Contract*, *tenure*, *OnlineSecurity*, *MonthlyCharges*, dan *TechSupport*. Strategi retensi yang direkomendasikan adalah: (1) paket *entry-level* berkontrak satu tahun dengan tarif tetap dan biaya aktivasi rendah, (2) *cross-sell* layanan internet dasar (DSL) berbiaya rendah untuk meningkatkan keterikatan layanan, (3) uji coba gratis *OnlineSecurity* selama 30–60 hari sebagai jalur konversi, (4) program *referral* berbasis kuota atau potongan tarif, serta (5) komunikasi retensi berbiaya rendah melalui SMS dan aplikasi *self-service*.

Strategi tersebut disusun atas dasar prinsip kesesuaian antara akar penyebab churn yang teridentifikasi melalui SHAP dengan instrumen intervensi yang tersedia bagi operator. Pada Cluster 0, intervensi diarahkan pada penguatan pengikat kontraktual dan mitigasi sensitivitas harga selama periode awal berlangganan. Pada Cluster 1, intervensi difokuskan pada perlindungan nilai bagi pelanggan bernilai tinggi melalui penguncian tarif dan peningkatan kualitas layanan premium, sebab tekanan *MonthlyCharges* tertinggi berpotensi mengikis loyalitas yang telah terbangun. Pada Cluster 2, intervensi ditekankan pada konversi pelanggan ekonomis menjadi pelanggan berkontrak melalui paket bertarif rendah dan strategi *cross-sell*.

Pendekatan retensi tersegmentasi ini secara teoretis memberikan rasio biaya-manfaat yang lebih optimal dibandingkan strategi seragam, karena alokasi anggaran retensi diselaraskan dengan kontribusi pendapatan, profil risiko, dan elastisitas respons masing-masing segmen. Pelanggan Cluster 1 yang menyumbang *TotalCharges* tertinggi diprioritaskan dengan instrumen retensi bernilai tinggi, sementara Cluster 2 yang berkontribusi pendapatan rendah ditangani melalui kanal retensi berbiaya rendah. Kerangka ini sejalan dengan rekomendasi Wu et al. (Wu et al., 2021) dan Chang et al. (Chang et al., 2024) yang menekankan bahwa interpretabilitas model harus diterjemahkan menjadi tindakan bisnis yang spesifik per segmen agar memberikan dampak retensi yang terukur.

5. KESIMPULAN

Penelitian ini berhasil mengusulkan dan mengimplementasikan kerangka analitik terintegrasi untuk prediksi churn pelanggan telekomunikasi yang menggabungkan prediksi *machine learning*, segmentasi K-Means, dan analisis interpretabilitas berbasis SHAP secara kohesif. Random Forest dengan SMOTE memberikan performa terbaik (Accuracy 0.7763, F1-Score 0.6114, ROC-AUC 0.833) dan terbukti meningkatkan kemampuan model dalam mengidentifikasi pelanggan churn (*Recall*) yang merupakan metrik kritis dalam konteks retensi. Segmentasi K-Means menghasilkan tiga cluster dengan karakteristik distinctif: Pelanggan Baru Rentan Pergi ($n = 1.325$), Pelanggan Loyal Bernilai Tinggi ($n = 462$), dan Pelanggan Hemat ($n = 474$).

Temuan paling signifikan diperoleh dari analisis SHAP per segmen yang mengungkap heterogenitas determinan churn antar cluster: fitur *Contract* konsisten menempati peringkat pertama, namun urutan fitur berikutnya berbeda secara substantif. *Tenure* menonjol pada Cluster 0 dan Cluster 2 sebagai indikasi kerentanan pelanggan baru, sedangkan *MonthlyCharges* mendominasi Cluster 1 yang mencerminkan sensitivitas pelanggan bernilai tinggi terhadap tekanan tarif. Heterogenitas ini menegaskan bahwa interpretasi global tidak memadai untuk merumuskan kebijakan retensi yang presisi, sehingga kontribusi utama penelitian terletak pada integrasi prediksi–segmentasi–interpretabilitas dalam satu *pipeline* serta penyajian SHAP per segmen sebagai dasar perumusan strategi retensi spesifik yang dapat ditindaklanjuti operator.

Penelitian ini memiliki keterbatasan pada penggunaan dataset publik IBM Telco, *Silhouette Score* 0.2819 yang tergolong moderat akibat dominasi fitur kategorikal, serta strategi retensi yang masih bersifat analitis dan belum divalidasi melalui implementasi lapangan. Penelitian lanjutan disarankan untuk: (1) menerapkan kerangka ini pada data operator nyata guna menguji *robustness* dan generalisabilitas temuan, (2) mengeksplorasi algoritma clustering alternatif seperti DBSCAN, *Gaussian Mixture Model*, atau *Hierarchical Clustering*, (3) mengintegrasikan algoritma *boosting* modern (LightGBM, CatBoost) dan *feature engineering* lanjutan untuk meningkatkan F1-Score, (4) memvalidasi efektivitas strategi retensi tersegmentasi melalui eksperimen *A/B testing*, serta (5) menambahkan dimensi *cost-sensitive learning* dan analisis *time-series* untuk menangkap dinamika perilaku pelanggan dari waktu ke waktu.

DAFTAR REFERENSI

- Adeniran, I. A., Efunniyi, C. P., Osundare, O. S., & Abhulimen, A. O. (2024). Implementing machine learning techniques for customer retention and churn prediction in telecommunications. *Computer Science & IT Research Journal*, 5(8), 2011–2025. <https://doi.org/10.51594/csitrj.v5i8.1489>
- Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, 34(4), 1582–1602. <https://doi.org/10.1287/isre.2023.1199>
- Chang, V., Hall, K., Xu, Q., Amao, F., Ganatra, M., & Benson, V. (2024). Prediction of customer churn behavior in the telecommunication industry using machine learning models. *Algorithms*, 17(6), Article 231. <https://doi.org/10.3390/a17060231>
- Esfahani, H. H., Toyonaga, S., & Oyibo, K. (2025). The application of explainable artificial intelligence in the prediction, diagnoses, treatment, and management of chronic diseases: A systematic review. *Digital Health*, 11. <https://doi.org/10.1177/20552076251355669>
- Harani, N. H., Prianto, C., & Nugraha, F. A. (2020). Segmentasi pelanggan produk digital service IndiHome menggunakan algoritma K-Means berbasis Python. *Jurnal Manajemen Informatika (JAMIKA)*, 10(2), 133–146. <https://doi.org/10.34010/jamika.v10i2.2683>
- IBM. (2018). *Telco customer churn dataset* [Data set]. Kaggle. <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>
- Imani, M., Ghaderpour, Z., Joudaki, M., & Beikmohammadi, A. (2024). *The impact of SMOTE and ADASYN on random forest and advanced gradient boosting techniques in telecom customer churn prediction*. Preprints. <https://doi.org/10.20944/preprints202403.0213.v2>
- Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. *Computing*, 104(2), 271–294. <https://doi.org/10.1007/s00607-021-00908-y>
- Leung, C. K., Pazdor, A. G. M., & Souza, J. (2021). Explainable artificial intelligence for data science on customer churn. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1–10). IEEE. <https://doi.org/10.1109/DSAA53316.2021.9564166>
- Mahmud, S., Islam, B. U., Anik, N. H., & Ghosh, T. (2024). Diabetes prediction: A comparative analysis of machine learning algorithms with SMOTE. In *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)* (pp. 1–6). IEEE. <https://doi.org/10.1109/COMPAS60761.2024.10796405>
- Noviandy, T. R., Idroes, G. M., Hardi, I., Afjal, M., & Ray, S. (2024). A model-agnostic interpretability approach to predicting customer churn in the telecommunications industry. *Infolitika Journal of Data Science*, 2(1), 34–44. <https://doi.org/10.60084/ijds.v2i1.199>
- Onoja, J. P. (2025). *A comprehensive review of customer churn prediction models in telecommunications*.

- Özkurt, C. (2025). Comparative analysis of XAI techniques on telecom churn prediction using SHAP and interpreted ML partial dependence. *Türk Doğa ve Fen Dergisi*, 14(2), 11–25. <https://doi.org/10.46810/tdfd.1529139>
- Salehi, A. R., & Khedmati, M. (2024). A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data. *Scientific Reports*, 14(1), Article 5152. <https://doi.org/10.1038/s41598-024-55598-1>
- Sikri, A., Jameel, R., Idrees, S. M., & Kaur, H. (2024). Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-63750-0>
- Wu, S., Yau, W.-C., Ong, T.-S., & Chong, S.-C. (2021). Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access*, 9, 62118–62136. <https://doi.org/10.1109/ACCESS.2021.3073776>