JAMBURA JOURNAL OF INFORMATICS
Volume 8, No. 2 | October 2026 : 132-146
ISSN 2685-4244, doi: 10.37905/jji.v8i2.41533

Integrasi *K-Means* dan *decision tree* dalam menganalisis kerentanan pengguna internet terhadap kejahatan siber

Rizki Nabillah, Agustiena Merdekawati, Azis Sukma Dhiana

Sistem Informasi, Universitas Bina Sarana Informatika, Jakarta, Indonesia

Riwayat Artikel:
Diterima 14 September 2026
Revisi 21 September 2026
Disetujui 10 Oktober 2026

Kata Kunci:
Kejahatan siber
Keamanan digital
Perilaku pengguna
K-Means clustering
Decision tree C4.5

ABSTRACT. Increasing internet use has heightened exposure to cybercrime, highlighting the need to identify users' vulnerability profiles and distinguishing characteristics. This study integrates K-Means Clustering and Decision Tree C4.5 using questionnaire data from 100 active internet users. K-Means grouped respondents by risk levels and security behaviors into four clusters, selected using the Elbow Method and Davies-Bouldin Index. Cluster labels were subsequently classified using independent predictors, including knowledge, perception, security awareness, age, and gender, to prevent circularity bias. SMOTE was applied exclusively to training data to address class imbalance without data leakage. The model achieved 56.00% accuracy and a macro F1-score of 51.44%. SMOTE increased recall for the low-vulnerability minority class from 30.00% to 40.00%, with 80.00% precision. Cybercrime perception emerged as the primary decision-tree splitting attribute. These findings reveal a knowledge–behavior gap and provide exploratory decision rules to inform behavior-based cybersecurity education.

ABSTRAK. Meningkatnya penggunaan internet turut memperbesar risiko kejahatan siber sehingga diperlukan pemetaan kerentanan pengguna berdasarkan karakteristik dan perilaku keamanan digital. Penelitian ini bertujuan mengidentifikasi profil kerentanan pengguna internet serta faktor yang membedakan setiap kelompok melalui integrasi K-Means Clustering dan Decision Tree C4.5. Data diperoleh melalui kuesioner terhadap 100 pengguna internet aktif. K-Means digunakan untuk mengelompokkan responden berdasarkan tingkat risiko dan perilaku keamanan, dengan jumlah klaster optimal ditentukan menggunakan Elbow Method dan Davies-Bouldin Index. Selanjutnya, hasil klasterisasi dijadikan label target pada Decision Tree C4.5 dengan memisahkan variabel prediktor berupa pengetahuan, persepsi, kesadaran keamanan, usia, dan jenis kelamin untuk menghindari *circularity bias*. Ketidakseimbangan kelas ditangani menggunakan SMOTE yang diterapkan hanya pada data latih guna mencegah *data leakage*. Pengujian menghasilkan akurasi 56,00% dan *macro F1-score* 51,44%. Penerapan SMOTE meningkatkan *recall* kelompok kerentanan rendah dari 30,00% menjadi 40,00%, dengan presisi 80,00%. Persepsi terhadap kejahatan siber teridentifikasi sebagai atribut utama dalam pembentukan pohon keputusan. Temuan ini mengungkap kesenjangan antara pengetahuan dan perilaku keamanan (*knowledge-behavior gap*) serta menghasilkan profil kerentanan dan aturan keputusan yang dapat menjadi dasar pengembangan edukasi keamanan siber berbasis perilaku.

This is an open-access article under the CC–BY-SA license

Penulis Korespondensi:
Agustiena Merdekawati,
Sistem Informasi, Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta, Indonesia
Email: merdekawatiagustiena@gmail.com

PENDAHULUAN

Internet telah menjadi bagian integral dalam kehidupan masyarakat dan digunakan untuk berbagai aktivitas, mulai dari komunikasi, pendidikan, pekerjaan, perdagangan hingga layanan pemerintahan. Berdasarkan data International Telecommunication Union (ITU), pada tahun 2026 sekitar 74% populasi dunia atau sekitar 6 miliar orang telah menggunakan internet. Peningkatan penggunaan internet memberikan berbagai kemudahan dan efisiensi, tetapi juga memperluas potensi risiko keamanan digital. Salah satu permasalahan yang berkembang adalah kejahatan siber (cybercrime), yaitu tindakan kriminal yang memanfaatkan komputer, jaringan internet, atau sistem elektronik sebagai alat, media, maupun objek kejahatan. Bentuknya meliputi phishing, malware, hacking, pencurian identitas, dan penyebaran konten ilegal yang dapat menyebabkan pencurian informasi, kerugian ekonomi, gangguan sistem, serta pelanggaran privasi.

Kondisi tersebut juga terlihat di Indonesia. Laporan ID-SIRTII/CC tahun 2025 mencatat 425.942.679 anomali trafik pada Januari 2025 dan menempatkan Indonesia sebagai negara sumber anomali tertinggi dalam kelompok lima negara yang dilaporkan. Tingginya aktivitas ancaman tersebut menunjukkan bahwa keamanan siber menjadi persoalan yang perlu diperhatikan. Namun, data insiden secara umum belum menjelaskan siapa pengguna yang memiliki tingkat kerentanan lebih tinggi dan karakteristik apa yang membedakan pengguna tersebut. Pada tingkat individu, kerentanan dapat berkaitan dengan perilaku digital, pengalaman menjadi korban, pengetahuan keamanan siber, perilaku keamanan, persepsi terhadap kejahatan siber, dan kesadaran keamanan. Oleh karena itu, permasalahan penelitian tidak hanya terletak pada peningkatan kejahatan siber, tetapi pada kebutuhan untuk mengidentifikasi kelompok pengguna berdasarkan tingkat kerentanan serta mengetahui atribut yang membedakan kelompok tersebut. Penelitian ini menggunakan data primer dari 100 responden yang mencakup karakteristik penggunaan internet, pengalaman kejahatan siber, pengetahuan, dan perilaku keamanan. Untuk mengidentifikasi pola tersebut, K-Means digunakan untuk mengelompokkan responden berdasarkan kemiripan karakteristik tanpa memerlukan label kelas sebelumnya.

Namun, K-Means hanya menghasilkan kelompok dan belum menjelaskan atribut yang membedakan setiap kelompok. Keterbatasan tersebut menjadi dasar penggunaan Decision Tree. Hasil clustering digunakan sebagai kelas untuk Decision Tree sehingga dapat diidentifikasi atribut yang paling berperan dalam membedakan kelompok kerentanan melalui ukuran *information gain*. Dengan demikian, K-Means menjawab bagaimana pengguna dikelompokkan, sedangkan Decision Tree membantu menjelaskan atribut apa yang membedakan kelompok tersebut.

Research gap penelitian ini terlihat dari beberapa penelitian terdahulu. Daeng *et al*. (2023) membahas keamanan siber di Indonesia dan menunjukkan pentingnya faktor manusia serta kesadaran keamanan digital, tetapi fokusnya masih pada kondisi keamanan siber secara umum dan belum mengelompokkan pengguna berdasarkan tingkat kerentanan. Ariyaningsih *et al*. (2023) membahas hubungan percepatan digitalisasi dengan kejahatan siber, tetapi menggunakan pendekatan deskriptif dan data sekunder sehingga belum menggambarkan pola kerentanan berdasarkan data primer pengguna. Sementara itu, Muningsih (2022) telah mengkombinasikan K-Means dan Decision Tree, tetapi objek penelitiannya berbeda, yaitu 811 data dengan 52 atribut, sehingga belum secara khusus menganalisis kerentanan pengguna internet terhadap kejahatan siber.

Research gap juga terdapat pada penentuan jumlah cluster. Secara metodologis, sebagian besar studi terdahulu yang mengombinasikan algoritma klasterisasi dan klasifikasi, seperti Muningsih (2022) serta Sari *et al*. (2025), masih menitikberatkan perbedaan pada ukuran dataset maupun variasi objek amatan. Pendekatan tersebut umumnya memperlakukan hasil klasterisasi sekadar sebagai pembuat label parsial tanpa melakukan evaluasi ketat terhadap potensi *data leakage* dan *class imbalance* saat ditransisikan ke tahap pemodelan *supervised learning*.

Penelitian ini menawarkan kontribusi dan kebaruan metodologis (*methodological novelty*) melalui perancangan kerangka kerja integrasi *unsupervised-supervised learning* yang terstruktur rapat. Kebaruan tersebut diwujudkan melalui:

1. Validasi jumlah klaster optimal berbasis *dual-metric evaluation* (Elbow Method dan Davies-Bouldin Index) sebelum penentuan variabel target,
2. Penerapan *strict feature isolation* antara variabel pembentuk klaster (X_1 dan X_3) dengan prediktor *Decision Tree* (X_2, X_4, X_5, dan demografi) untuk memutuskaitkan bias sirkularitas (*circularity bias*), serta
3. Integrasi teknik *Synthetic Minority Over-sampling Technique* (SMOTE) khusus pada data latih guna mengatasi ketidakseimbangan kelas (*class imbalance*) hasil *clustering*.

Melalui kerangka kerja ini, analisis kerentanan keamanan siber tidak hanya menghasilkan pemetaan kelompok berbasis data primer, tetapi juga menjamin keandalan aturan keputusan (*decision rules*) yang mengekstrak fenomena *knowledge-behavior gap* secara objektif (Fernández-Navarro *et al*., 2023a)."

Berdasarkan gap tersebut, penelitian ini menawarkan integrasi K-Means Clustering dan Decision Tree untuk menganalisis kerentanan pengguna internet terhadap kejahatan siber. K-Means digunakan untuk menemukan kelompok pengguna berdasarkan kemiripan karakteristik, sedangkan hasil cluster digunakan sebagai kelas pada Decision Tree untuk mengidentifikasi atribut pembeda dan menghasilkan aturan keputusan.

Novelty penelitian terletak pada penerapan integrasi unsupervised learning dan supervised learning pada analisis kerentanan pengguna internet berbasis data primer, dengan penentuan jumlah cluster melalui pengujian beberapa nilai *k* menggunakan Elbow dan Davies-Bouldin Index sebelum dilakukan analisis Decision Tree. Dengan pendekatan tersebut, penelitian tidak hanya menghasilkan pengelompokan tingkat kerentanan, tetapi juga memberikan interpretasi mengenai karakteristik yang membedakan kelompok pengguna.

Berdasarkan uraian tersebut, penelitian ini bertujuan menganalisis kerentanan pengguna internet terhadap kejahatan siber melalui integrasi K-Means Clustering dan Decision Tree untuk mengidentifikasi pola kelompok pengguna serta atribut yang paling berperan dalam membedakan tingkat kerentanan.

METODE

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis pola kerentanan pengguna internet terhadap kejahatan siber berdasarkan perilaku dan karakteristik keamanan digital. Data yang digunakan berupa data primer yang diperoleh melalui kuesioner kepada pengguna internet aktif.

Kerangka penelitian (Syafrizal *et al*., 2025) adalah konsep pada penelitian yang saling berhubungan, dimana penggambaran variabel satu dengan yang lainnya bisa terkoneksi secara detail dan sistematis. Gambar 1 merupakan alur penelitian yang dilakukan.



Gambar 1. Alur penelitian

Integrasi kedua metode dilakukan secara bertahap. *K-Means* digunakan terlebih dahulu untuk menemukan kelompok pengguna berdasarkan tingkat kerentanan, sedangkan *Decision Tree* digunakan pada tahap berikutnya untuk mengidentifikasi karakteristik pembeda antar kelompok yang telah terbentuk.

Variabel dan Instrumen Penelitian

Variabel penelitian ditentukan berdasarkan kajian literatur mengenai keamanan digital dan perilaku pengguna internet. Variabel yang digunakan untuk membentuk profil kerentanan terdiri atas lima konstruk, yaitu Tingkat Risiko (X1), Pengetahuan terhadap Kejahatan Siber (X2), Perilaku Keamanan (X3), Persepsi terhadap Kejahatan Siber (X4), dan Kesadaran Keamanan (X5).

Setiap konstruk diukur menggunakan beberapa indikator yang dituangkan dalam instrumen kuesioner. Setiap butir pernyataan menggunakan Skala Likert lima tingkat, yaitu 1 = Sangat Tidak Setuju, 2 = Tidak Setuju, 3 = Netral, 4 = Setuju, dan 5 = Sangat Setuju. Instrumen diuji validitas dan reliabilitasnya sebelum digunakan dalam proses analisis.

Variabel pembentuk klaster terdiri dari 2 variabel konstruktor, yaitu variable tingkat resiko (X1) dan perilaku keamanan (X3). Untuk pengolahan decision tree menggunakan variabel lainnya, seperti pengetahuan kejahatan siber (X2), persepsi kejahatan siber (X4), kesadaran keamanan (X5), jenis kelamin dan usia.

Populasi dan Sampel

Populasi penelitian adalah pengguna internet yang aktif menggunakan internet dalam aktivitas sehari-hari. Sampel ditentukan menggunakan teknik *purposive sampling* dengan kriteria responden merupakan pengguna internet aktif dan bersedia mengisi kuesioner penelitian. Di mana responden dipilih berdasarkan kriteria inklusi spesifik: (1) pengguna internet aktif dengan durasi harian $\geq$ 4 jam, dan (2) pernah melakukan transaksi atau aktivitas digital interaktif. Penggunaan teknik ini bertujuan untuk memastikan bahwa data yang diperoleh berasal dari individu yang relevan dengan profil kerentanan kejahatan siber.

Sebanyak 131 responden berhasil dikumpulkan melalui penyebaran kuesioner secara daring menggunakan Google Forms yang didistribusikan melalui WhatsApp. Jumlah tersebut digunakan sebagai keseluruhan data penelitian yang memenuhi kriteria dan tidak menggunakan perhitungan Slovin sebagai dasar penentuan jumlah sampel. Dengan demikian, penggunaan *purposive sampling* tetap konsisten dengan karakteristik penelitian yang memilih responden berdasarkan kriteria tertentu, sehingga generalisasi hasil terbatas pada karakteristik populasi serupa. Penelitian selanjutnya disarankan untuk memperluas cakupan sampel dan menggunakan teknik probability sampling untuk meningkatkan daya generalisasi model.

Pengumpulan dan Pengujian Instrumen

Data primer dikumpulkan melalui kuesioner daring. Jawaban responden kemudian diekspor ke dalam Microsoft Excel untuk dilakukan pemeriksaan awal, pengkodean, dan penyusunan dataset.

Sebelum digunakan dalam analisis *machine learning*, instrumen diuji validitas dan reliabilitasnya menggunakan IBM SPSS Statistics. Uji validitas digunakan untuk mengetahui kemampuan setiap butir pernyataan dalam mengukur konstruk yang diteliti, sedangkan uji reliabilitas menggunakan *Cronbach's Alpha* untuk menilai konsistensi internal instrumen. Butir yang tidak memenuhi kriteria pengujian ditinjau sebelum data digunakan pada tahap analisis berikutnya.

Suatu instrumen dapat dikatakan reliabel apabila memiliki nilai *Cronbach's Alpha* > **0,60**, yang menunjukkan bahwa suatu item pertanyaan memiliki tingkat kestabilan dan dapat menghasilkan data yang konsisten.

Preprocessing Data

Tahap *preprocessing* dilakukan untuk menghasilkan dataset yang bersih dan sesuai dengan kebutuhan algoritma. Proses diawali dengan *editing* terhadap kelengkapan jawaban, dilanjutkan dengan pemeriksaan data duplikat dan *missing value*. Data yang tidak memenuhi kriteria kelengkapan penelitian dikeluarkan dari dataset.

Jawaban Skala Likert kemudian dikonversi menjadi nilai numerik 1–5. Untuk setiap konstruk penelitian, skor indikator dihitung sesuai dengan struktur instrumen sehingga diperoleh nilai konstruk Tingkat Risiko, Pengetahuan terhadap Kejahatan Siber, Perilaku Keamanan, Persepsi terhadap Kejahatan Siber, dan Kesadaran Keamanan.

Data numerik yang digunakan pada proses *K-Means* selanjutnya dinormalisasi menggunakan metode *Min-Max Normalization* ke dalam rentang 0–1. Normalisasi dilakukan karena *K-Means* menggunakan perhitungan jarak sehingga perbedaan skala antarvariabel dapat memengaruhi pembentukan klister hasil normalisai *Min-Max* dapat dilihat pada Tabel 1.

Tabel 1. Hasil normalize

Nilai Asli	Hasil Normalize
1	0
2	0.25
3	0.50
4	0.75
5	1

Proses pengumpulan data awal berhasil memperoleh 131 responden. Namun, pada tahap seleksi dan pembersihan data (*data cleaning*), dilakukan penyaringan berbasis kriteria inklusi dan pemeriksaan kualitas isian. Sebanyak 31 data dikeluarkan dari analisis akhir karena memenuhi kriteria eksklusi, yaitu: (1) tidak memenuhi kriteria inklusi belum pernah mengalami kejahatan siber, (2) indikasi pengisian tidak serius (*uniform/straight-lining response* pada skala Likert), serta (3) adanya data duplikat (*missing values*) pada variabel utama. Dengan demikian, dataset bersih akhir yang dianalisis secara konsisten berjumlah 100 responden ($\boldsymbol{N = 100}$).

Pembentukan Klaster dengan K-Means

Tahap pertama machine learning menggunakan algoritma K-Means Clustering. Algoritma ini digunakan untuk mengelompokkan responden berdasarkan kemiripan profil kerentanan yang dibentuk dari dua konstruk penelitian, yaitu variabel tingkat risiko ($\boldsymbol{X_1}$) dan perilaku keamanan ($\boldsymbol{X_3}$). Secara umum, tujuan K-Means adalah meminimalkan jarak antara data dengan pusat klasternya melalui fungsi objective berikut (Sinaga & Yang, 2020):

$$\boldsymbol{J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2} \qquad (1)$$

Dengan $\boldsymbol{K}$ sebagai jumlah klaster, $\boldsymbol{C_k}$ sebagai kelompok ke-$\boldsymbol{k}$, $\boldsymbol{x_i}$ sebagai data responden, dan $\boldsymbol{\mu_k}$ sebagai centroid klaster.

Penentuan jumlah klaster dalam penelitian ini tidak dilakukan secara subjektif, melainkan melalui pengujian terhadap beberapa alternatif jumlah klaster. Berdasarkan hasil pengujian, rentang jumlah klaster yang diuji adalah $\boldsymbol{K = 2}$ hingga $\boldsymbol{K = 7}$. Evaluasi jumlah klaster dilakukan dengan mempertimbangkan Elbow Method dan Davies-Bouldin Index (DBI). Elbow Method digunakan untuk mengamati perubahan nilai Within-Cluster Sum of Squares (WCSS) pada setiap alternatif jumlah klaster (Kassambara, 2017), sedangkan DBI digunakan untuk mengevaluasi kualitas

pengelompokan berdasarkan kekompakan data dalam klaster dan pemisahan antar-klaster (Syakur *et al*., 2018).

Penentuan jumlah klaster optimal dilakukan secara komprehensif dengan mengombinasikan hasil pengujian Elbow Method dan Davies-Bouldin Index (DBI) pada rentang K=2 hingga K=7. Pemilihan jumlah klaster tidak hanya didasarkan pada skor matematis DBI semata, melainkan juga mempertimbangkan titik siku (elbow point) pada grafik Sum of Squared Errors (SSE), keharmonisan distribusi anggota klaster, serta pencegahan terbentuknya klaster pencilan (outlier cluster) agar kelompok yang dihasilkan memiliki nilai interpretasi domain yang kuat dalam menganalisis kerentanan pengguna internet.

Setelah proses pengelompokan selesai, setiap responden memperoleh label klaster yang selanjutnya digunakan sebagai variabel target pada tahap pembentukan model Decision Tree. Sebelum digunakan dalam analisis lanjutan, masing-masing klaster diinterpretasikan untuk memahami profil dan karakteristik kerentanan pengguna internet secara mendalam pada setiap kelompok (Prusty *et al*., 2022).

Analisis Decision Tree

Decision Tree digunakan setelah pembentukan klaster untuk mengidentifikasi karakteristik responden yang membedakan kelompok kerentanan yang telah diperoleh dari *K-Means*. Pada tahap ini, label klaster hasil *K-Means* digunakan sebagai variabel target.

Untuk menghindari *circularity* dan *data leakage*, variabel tingkat resiko (X1) dan perilaku keamanan (X3) yang digunakan untuk membentuk klaster, tidak digunakan kembali sebagai prediktor *Decision Tree*. Prediktor pada tahap ini berasal dari karakteristik responden yang tidak digunakan dalam pembentukan klaster, seperti kelompok usia, jenis kelamin, variabel pengetahuan kejahatan siber (X2), persepsi kejahatan siber, dan kesadaran keamanan (X5).

Pendekatan tersebut memungkinkan penelitian menjawab dua pertanyaan yang berbeda. *K-Means* digunakan untuk menemukan kelompok kerentanan berdasarkan profil keamanan digital, sedangkan *Decision Tree* digunakan untuk mengetahui karakteristik eksternal yang dapat membedakan kelompok tersebut.

Untuk menangani ketidakseimbangan kelas (*class imbalance*) hasil klasterisasi K-Means, teknik Synthetic Minority Over-sampling Technique (SMOTE) diterapkan. Guna mencegah terjadinya data leakage dan evaluasi yang bias, proses SMOTE diisolasi secara ketat dan hanya diaplikasikan pada data pelatihan (train set). Evaluasi kinerja model klasifikasi (confusion matrix, akurasi, presisi, recall, dan F1-score) dilakukan secara murni menggunakan data pengujian asli (test set) yang tidak mengalami oversampling.

Apabila terdapat empat klaster hasil *K-Means*, maka seluruh empat label tersebut harus dipertahankan sebagai kelas pada *Decision Tree*. Tidak dilakukan pengurangan kelas secara sepihak. Apabila salah satu klaster tidak dapat dipertahankan setelah pemeriksaan *outlier* dan evaluasi klaster, jumlah kelas yang digunakan pada *Decision Tree* harus disesuaikan dengan hasil akhir *K-Means*.

Analisis dan Interpretasi Hasil

Analisis dilakukan dengan menggabungkan hasil evaluasi *K-Means* dan *Decision Tree*. Hasil *K-Means* digunakan untuk menggambarkan struktur kelompok pengguna berdasarkan tingkat kerentanan, sedangkan hasil *Decision Tree* digunakan untuk menjelaskan karakteristik yang membedakan kelompok tersebut.

Interpretasi dilakukan dengan memperhatikan ukuran dan karakteristik setiap klaster, kualitas pemisahan klaster berdasarkan DBI dan *Elbow Method*, serta aturan klasifikasi yang dihasilkan oleh

Decision Tree. Hasil tersebut selanjutnya digunakan untuk merumuskan gambaran kelompok pengguna yang memiliki karakteristik kerentanan berbeda dan menjadi dasar penyusunan rekomendasi edukasi keamanan digital.

Dengan rancangan tersebut, hasil penelitian tidak dimaksudkan untuk menyatakan hubungan sebab-akibat antarvariabel, tetapi untuk mengidentifikasi pola pengelompokan dan karakteristik yang membedakan kelompok pengguna internet berdasarkan kerentanan terhadap kejahatan siber.

HASIL DAN DISKUSI

Profil dan Karakteristik Responden

Penelitian ini menggunakan data primer yang diperoleh melalui kuesioner dan menghasilkan sebanyak 100 responden. Karakteristik responden dianalisis berdasarkan jenis kelamin, kelompok usia, dan pengalaman menjadi korban kejahatan siber. Berdasarkan data yang diperoleh, responden perempuan berjumlah 59%, sedangkan responden laki-laki sebesar 41%. Sebagian besar responden berada pada kelompok usia 20–35 tahun, yaitu sebesar 87%. Selain itu, pengalaman responden dengan bentuk pengalaman yang paling banyak yaitu 45% responden menyatakan pernah mengalami kejahatan siber berupa penipuan online dan hoak sebesar 38%,.

Karakteristik tersebut memberikan gambaran awal mengenai profil responden sebelum dilakukan proses pengelompokan. Namun, karakteristik demografis tersebut tidak secara langsung digunakan dalam tahap pengelompokan awal. Sesuai dengan desain analisis yang menerapkan isolasi fitur secara ketat (*strict feature isolation*), pengelompokan K-Means dilakukan khusus berdasarkan dua konstruk profil keamanan digital, yaitu Tingkat Risiko (X_1) dan Perilaku Keamanan (X_3). Sementara itu, tiga konstruk lainnya (Pengetahuan Kejahatan Siber (X_2), Persepsi Kejahatan Siber (X_4), dan Kesadaran Keamanan (X_5)) beserta karakteristik demografi (Usia dan Jenis Kelamin) dipisahkan secara konsisten untuk digunakan sebagai variabel prediktor pembeda dalam pemodelan Decision Tree.

Hasil Uji Kualitas Data (Pengujian Instrumen)

Sebelum data primer diolah menggunakan algoritma *machine learning*, instrumen penelitian berbasis Skala Likert (1–5) diuji terlebih dahulu kualitasnya untuk menjamin tingkat ketepatan (*validitas*) dan keandalan (*reliabilitas*) data. Variabel yang diuji meliputi Tingkat Risiko (X_1), Pengetahuan Kejahatan Siber (X_2), Perilaku Keamanan (X_3), Persepsi Kejahatan Siber (X_4), dan Kesadaran Keamanan (X_5).

Pengujian validitas dilakukan dengan membandingkan nilai korelasi Pearson (r_{hitung}) terhadap r_{tabel} pada tingkat signifikansi 5% dengan jumlah sampel $N = 100$. Seluruh item pertanyaan untuk kelima variabel penelitian dinyatakan valid karena nilai r_{hitung} pada setiap indikator melampaui ambang batas r_{tabel} serta memiliki korelasi yang signifikan ($p < 0{,}05$). Berdasarkan ke lima variable yang diuji menggunakan validitas, memberikan hasil pengujian bahwa r Hitung > r Tabel. Berikut hasil uji validitas dari variabel resiko sebagai berikut:

Tabel 2. Uji validitas tingkat resiko

No	r Hitung	r Tabel	Keterangan
X1	0,399	0,1966	Valid
X2	0,686	0,1966	Valid
X3	0,632	0,1966	Valid
X4	0,657	0,1966	Valid
X5	0,673	0,1966	Valid

Selanjutnya, uji reliabilitas dilakukan dengan mengevaluasi nilai koefisien *Cronbach's Alpha*. Mengacu pada kriteria Guildford, instrumen penelitian dinyatakan reliabel dan memiliki kestabilan

yang tinggi untuk dipakai pada analisis tingkat lanjut jika nilai *Cronbach's Alpha* berada di atas kriteria minimal ($\geq 0{,}60$). Hasil pengujian menunjukkan bahwa instrumen kuesioner memiliki tingkat konsistensi internal yang kuat hingga sangat kuat, sehingga dataset layak digunakan untuk proses *clustering* dan *classification*, rincian hasil pengujian disajikan dalam Tabel 3.

Tabel 3. Hasil pengujian reliabilitas

Variabel	Cronbach Alpha	N of Items	Keterangan
Tingkat Resiko	0,772	4	Reliabel (Kuat)
Pengetahuan terhadap kejahatan siber	0,685	4	Reliabel (Kuat)
Perilaku Keamanan	0,699	6	Reliabel (Kuat)
Persepsi	0,698	5	Reliabel (Kuat)
Kesadaran Keamanan	0,665	4	Reliabel (Kuat)

Penentuan Jumlah Cluster

Dalam upaya mengelompokkan data responden berdasarkan tingkat kerentanan terhadap kejahatan siber, pengujian validitas klaster dilakukan terlebih dahulu untuk menentukan jumlah klaster (K) yang paling optimal. Evaluasi ini memanfaatkan dua indikator utama, yaitu *Davies-Bouldin Index* (DBI) dan *Average Within Centroid Distance* yang divisualisasikan menggunakan *Elbow Method*. Dalam penentuan jumlah klaster yang digunakan variable X_1 (Tingkat Resiko) dan X_3 (Perilaku Keamanan). Dalam menentukan jumlah klaster (K) yang paling optimal, dilakukan evaluasi menggunakan Davies-Bouldin Index (DBI) dan Elbow Method (Sum of Squared Errors / SSE) pada rentang K=2 hingga K=7. Pengujian dilakukan menggunakan variabel Tingkat Risiko (X1) dan Perilaku Keamanan (X3). Hasil pengujian DBI disajikan pada Tabel 4.

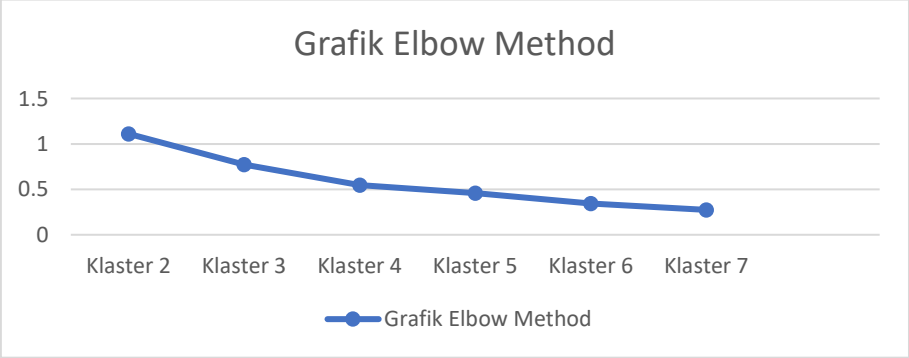
Tabel 4. Hasil pengujian DBI

Jumlah klaster	Nilai DBI
K = 2	0.977
K = 3	0.908
K = 4	0.825
K = 5	0.765
K = 6	0.756
K = 7	0.738

Berdasarkan Tabel 4, nilai DBI secara matematis terus mengalami penurunan hingga mencapai nilai minimum pada K=7 dengan skor sebesar 0,738. Namun, pengujian lebih lanjut terhadap pembagian anggota klaster pada K=7 menunjukkan terbentuknya outlier cluster, di mana salah satu klaster hanya beranggotakan 2 responden (2% dari total sampel), sementara klaster lainnya berkisar antara 10 hingga 26 responden. Keberadaan klaster mikro semacam ini tidak ideal untuk pemodelan supervised learning pada tahap berikutnya karena memicu extreme class imbalance.

Oleh karena itu, evaluasi diselaraskan dengan Elbow Method pada Gambar 2 dan Tabel 5 untuk melihat titik jenuh penurunan nilai SSE. Berdasarkan Tabel 5, terjadi penurunan SSE yang sangat signifikan dari K=2 (1,111) ke K=3 (0,774) dan K=4 (0,548). Saat jumlah klaster dinaikkan dari K=4 ke K=5 (0,460), penurunan nilai SSE mulai melandai secara drastis (hanya sebesar 0,088). Hal ini mengindikasikan bahwa penambahan klaster setelah K=4 tidak lagi memberikan peningkatan variansi antar-klaster yang signifikan.

Dengan mempertimbangkan kombinasi titik siku Elbow Method pada K=4, tren penurunan DBI yang konstan, serta terbentuknya distribusi keanggotaan yang seimbang tanpa adanya outlier cluster (Cluster 0 = 47, Cluster 1 = 10, Cluster 2 = 18, dan Cluster 3 = 25 data), maka jumlah klaster K=4 ditetapkan sebagai konfigurasi terbaik dan paling stabil untuk tahap analisis Decision Tree.



Gambar 2. Grafik Elbow methods

Tabel 5. Hasil *Elbow method*

Klaster	Elbow Method
Klaster 2	1.111
Klaster 3	0.774
Klaster 4	0.548
Klaster 5	0.46
Klaster 6	0.343
Klaster 7	0.274

Hasil Clustering K-Means

Berdasarkan jumlah cluster yang telah ditentukan, algoritma K-Means mengelompokkan 100 responden ke dalam 4 cluster. Distribusi anggota pada masing-masing cluster menunjukkan adanya perbedaan ukuran kelompok, yang mengindikasikan bahwa profil keamanan digital responden tidak tersebar secara merata. Distribusi anggota klaster yang diperoleh adalah sebagai berikut:

Tabel 6. Distribusi anggota klaster

Klaster	Jumlah Responden
Cluster 0	47
Cluster 1	10
Cluster 2	18
Cluster 3	25
Total	100

Berdasarkan hasil klasterisasi K-Means dengan $K = 4$, diperoleh distribusi jumlah anggota yang seimbang pada seluruh klaster. Dari total 100 responden, Cluster 0 mencakup 47 data (47%), Cluster 3 sebanyak 25 data (25%), Cluster 2 sebanyak 18 data (18%), dan Cluster 1 sebesar 10 data (10%). Keberadaan Cluster 1 dengan proporsi 10% menunjukkan bahwa tidak terdapat *outlier cluster* pada model $K = 4$. Pembagian ini menghasilkan representasi kelas target yang terdistribusi secara wajar tanpa adanya kecenderungan *extreme class imbalance*, sehingga sangat ideal dan siap digunakan sebagai variabel target (Y) pada tahap klasifikasi menggunakan algoritma *Decision Tree*.

Interpertasi Kerentanan Pengguna Internet

Berdasarkan hasil pengelompokan menggunakan algoritma *K-Means Clustering* terhadap 100 responden, terbentuk 4 kluster (*Cluster 0* hingga *Cluster 3*) yang memetakan tingkat kerentanan pengguna internet berdasarkan variabel tingkat resiko (X_1) dan tingkat perilaku keamanan (X_3). Gambaran umum pusat kluster (*centroid*) serta distribusi demografi responden disajikan pada Tabel 7.

Tabel 7. Pusat klaster (centroid) dan distribusi pengelompokkan kerentanan pengguna internet

Kluster	Jumlah (n)	Persentase	X1	X3	Dominasi Gender	Karakteristik Kerentanan
Cluster 0	47	47%	3,41	3,99	Perempuan (59,6%)	Moderat / Kerentanan Sedang-Tinggi
Cluster 1	10	10%	3,23	2,98	Seimbang (50% : 50%)	Kerentanan Rendah (*Low Risk*)
Cluster 2	18	18%	3,43	4,76	Perempuan (72,2%)	Kerentanan Sangat Tinggi (*Behavioral Gap*)
Cluster 3	25	25%	1,75	4,53	Seimbang (52% P : 48% L)	Kerentanan Kritis (*High Risk)*

Cluster 0: Kelompok Kerentanan Moderat/Kerentanan Sangat Tinggi (Moderate Vulnerability)

Cluster 0 merupakan kelompok terbesar yang mencakup 47% dari total sampel (47 responden; 28 perempuan dan 19 laki-laki, didominasi kelompok usia kategori 2). Kluster ini memiliki skor rata-rata Tingkat resiko (X_1) sebesar 3,41 dan skor tingkat perilaku keamanan (X_3) sebesar 3,99. Pengguna dalam kelompok ini memiliki tingkat pengetahuan dasar yang mencukupi terkait risiko internet, namun tetap berada dalam kategori kerentanan sedang hingga tinggi karena aktivitas digital yang intensif belum sepenuhnya diimbangi oleh tindakan proteksi yang konsisten.

Cluster 1: Kelompok Kerentanan Rendah (Low Risk / Resilient Users)

Cluster 1 merupakan kelompok terkecil dengan proporsi 10% (10 responden; 5 laki-laki dan 5 perempuan). Rata-rata nilai X_1 berada pada angka 3,23, sedangkan tingkat perilaku keamanan (X_3) tercatat paling rendah di antara seluruh kluster, yaitu 2,98. Pengguna pada kelompok ini tergolong paling aman dari ancaman siber karena mampu membatasi paparan risiko digital dan menerapkan kebiasaan berinternet yang relatif tertib.

Cluster 2: Kelompok Kesenjangan Perilaku (Behavioral Gap)/Kerentanan Sangat Tinggi

Cluster 2 mencakup 18% responden (18 orang; didominasi oleh perempuan sebanyak 72,2%). Kelompok ini menunjukkan fenomena yang cukup krusial: tingkat resiko (X_1) tergolong baik dengan nilai 3,43, namun skor tingkat perilaku keamanan (X_3) mencapai nilai tertinggi dibanding kluster lainnya, yaitu 4,76. Kondisi ini mengindikasikan adanya *knowledge-behavior gap*, di mana pemahaman teori keamanan siber yang dimiliki responden tidak langsung diimplementasikan dalam perilaku nyata sehari-hari, sehingga mereka tetap sangat rentan terhadap serangan digital.

Cluster 3: Kelompok Kerentanan Kritis (High-Risk / High Vulnerability)

Cluster 3 mencakup 25% dari total responden (25 orang; 13 perempuan dan 12 laki-laki). Kluster ini merupakan kelompok paling rawan karena memiliki nilai rata-rata X_1 paling rendah (1,75) yang berbanding lurus dengan nilai tingkat perilaku keamanan (X_3) yang sangat tinggi (4,53). Kombinasi antara minimnya pemahaman dasar mengenai keamanan siber dan tingginya tingkat paparan ancaman menjadikan *Cluster 3* sebagai sasaran utama yang paling membutuhkan intervensi berupa edukasi dan pelatihan literasi digital dasar.

Sintesis dari keempat kluster menunjukkan bahwa mayoritas pengguna internet (82% yang tersebar di *Cluster 0*, *Cluster 2*, dan *Cluster 3*) memiliki tingkat kerentanan pada kategori tinggi hingga sangat

tinggi ($X_3 \geq 3{,}99$). Tingkat pengetahuan yang memadai tidak secara otomatis menjamin keamanan pengguna jika tidak diiringi dengan kesadaran menerapkan protokol keamanan secara aktif. Prioritas edukasi dan pendampingan keamanan siber perlu difokuskan pada *Cluster 3* untuk peningkatan pemahaman dasar dan *Cluster 2* untuk mendorong perubahan perilaku protektif.

Hasil Klasifikasi Decision Tree

Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor serta pola kebiasaan pengguna internet yang paling signifikan memengaruhi tingkat kerentanan mereka terhadap kejahatan siber. Proses pengolahan data pada penelitian ini menerapkan kerangka kerja *Knowledge Discovery in Databases* (KDD) yang mencakup empat tahapan utama:

Data Selection (Seleksi Data)

Pada tahap data selection, dataset berukuran 100 responden dipilah untuk menentukan atribut masukan (prediktor) dan atribut target. Atribut prediktor yang dipilih meliputi variabel jenis kelamin, usia, pengetahuan kejahatan siber (X_2), persepsi mengenai kejahatan siber (X_4), serta kesadaran keamanan (X_5). Sementara itu, label kelompok hasil klasterisasi K-Means bertindak sebagai variabel target (Y) yang merepresentasikan empat kategori tingkat kerentanan (Kerentanan Rendah, Kerentanan Sedang/Tinggi, Kerentanan Kritis, dan Kerentanan Sangat Tinggi).

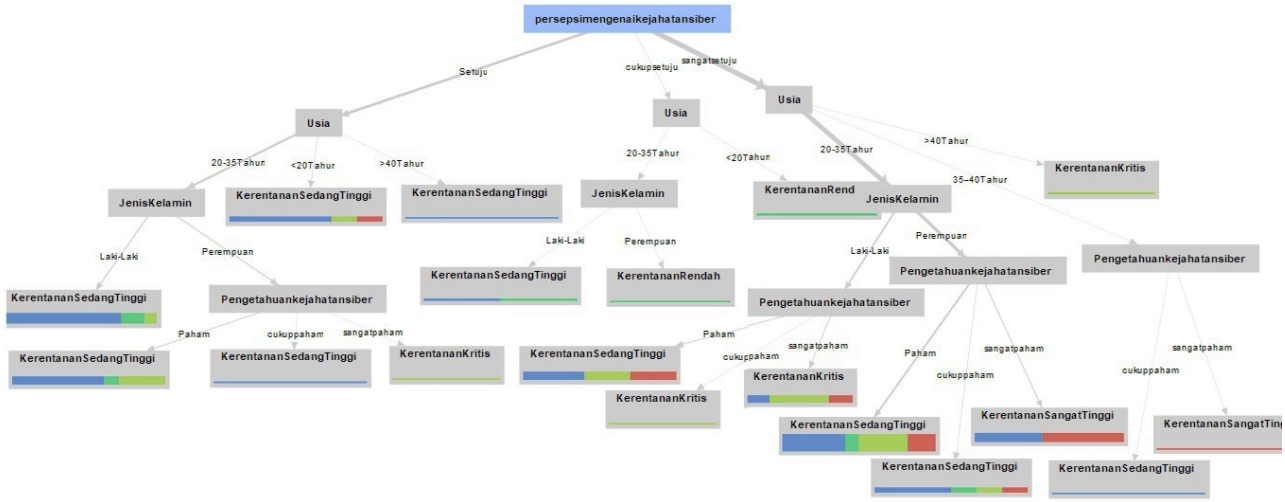
Preprocessing Data (Pra-pemrosesan Data)

Pada tahap *preprocessing*, dilakukan pembersihan data berupa penanganan spasi tersembunyi (*whitespace*) pada nilai kategori menggunakan teknik *trimming* dan *mapping* guna menghindari kesalahan pembacaan atribut pada pustaka pemodelan. Selanjutnya, analisis *Weight by Information Gain* diterapkan untuk mengukur tingkat kontribusi atribut, di mana variabel persepsi mengenai kejahatan siber terpilih sebagai *root node* atau simpul utama pembagi data. Untuk mengatasi permasalahan ketidakseimbangan kelas (*class imbalance*) di mana kelas Kerentanan Sedang/Tinggi sangat mendominasi dataset awal, digunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mereplikasi sampel kelas minoritas secara sintetis pada data latih.

Modelling (Pemodelan)

Pemodelan yang dilakukan menggunakan algoritma *Decision Tree* menghasilkan struktur pohon keputusan dan aturan (*rules*) klasifikasi. Berdasarkan struktur *hierarchy tree* setelah penerapan SMOTE, proses klasifikasi diawali oleh percabangan variabel persepsi mengenai kejahatan siber. Pada kelompok responden yang memiliki persepsi "Setuju", pola klasifikasi didominasi oleh kelompok usia 20–35 tahun. Responden laki-laki pada rentang usia tersebut secara mayoritas diklasifikasikan ke dalam kategori Kerentanan Sedang/Tinggi. Bagi responden perempuan, tingkat pengetahuan kejahatan siber menjadi faktor pembeda selanjutnya, di mana pengetahuan kategori "Paham" dan "Cukup Paham" mengarah pada Kerentanan Sedang/Tinggi, sedangkan pengetahuan "Sangat Paham" diklasifikasikan ke dalam kategori Kerentanan Kritis. Untuk kelompok usia kurang dari 20 tahun dan lebih dari 40 tahun dengan persepsi "Setuju", model secara konsisten memprediksi tingkat Kerentanan Sedang/Tinggi.

Sementara itu, pada cabang responden dengan persepsi "Cukup Setuju", kecenderungan tingkat kerentanan berada pada kategori yang lebih rendah. Responden berusia 20–35 tahun dengan jenis kelamin perempuan serta responden berusia kurang dari 20 tahun diklasifikasikan ke dalam kategori Kerentanan Rendah. Adapun pada kelompok responden dengan persepsi "Sangat Setuju", struktur aturan keputusan menjadi lebih kompleks dengan melibatkan variabel pengetahuan kejahatan siber. Pada kelompok usia 20–35 tahun, responden laki-laki dengan pengetahuan "Sangat Paham" diklasifikasikan sebagai Kerentanan Kritis, sedangkan responden perempuan dengan tingkat pengetahuan yang sama cenderung diprediksi ke dalam kategori Kerentanan Sangat Tinggi. Pada kelompok usia 35–40 tahun dan lebih dari 40 tahun dengan persepsi "Sangat Setuju", klasifikasi masing-masing didominasi oleh kelas Kerentanan Sangat Tinggi dan Kerentanan Kritis.



Gambar 3. Pohon keputusan

Evaluasi

Evaluasi kinerja model Decision Tree dilakukan dengan membandingkan performa model sebelum penyeimbangan data dan sesudah penerapan SMOTE. Sesuai dengan kaidah metodologi pembelajaran mesin untuk mencegah *data leakage*, proses *oversampling* menggunakan SMOTE hanya diterapkan pada data pelatihan (training set). Evaluasi performa model secara keseluruhan diuji menggunakan data pengujian asli (testing set murni, N=100) hasil pembagian rasio 80:20 tanpa perlakuan *oversampling*.

accuracy: 56.00%

	true KerentananSed...	true KerentananRen...	true KerentananKritis	true KerentananSan...	class precision
pred. KerentananSe...	35	6	14	7	56.45%
pred. KerentananRe...	1	4	0	0	80.00%
pred. KerentananKrit...	6	0	11	5	50.00%
pred. KerentananSa...	5	0	0	6	54.55%
class recall	74.47%	40.00%	44.00%	33.33%	

Gambar 4. Hasil *confussion matrix*

Berdasarkan Gambar 4, Kinerja Model Awal (Tanpa SMOTE): Model awal tanpa penyeimbangan data menghasilkan akurasi keseluruhan sebesar 59,00%. Namun, model ini mengalami bias pada kelas mayoritas (*Kerentanan Sedang/Tinggi*), ditunjukkan oleh *recall* yang tinggi pada kelas tersebut (85,11%). Sebaliknya, daya deteksi terhadap kelas minoritas sangat rendah, terutama pada *Kerentanan Rendah* dengan *recall* hanya 30,00% (hanya 3 dari 10 data aktual terdeteksi), serta *Kerentanan Kritis* (36,00%) dan *Kerentanan Sangat Tinggi* (38,89%). Setelah SMOTE diaplikasikan pada data pelatihan dan model diuji kembali pada *test set* murni ($N = 100$), diperoleh nilai akurasi sebesar 56,00%. Penerapan SMOTE pada *train set* berhasil memperbaiki sensitivitas model terhadap kelas minoritas:

- Kerentanan Rendah: *Recall* meningkat dari 30,00% menjadi 40,00% (berhasil mengklasifikasikan 4 dari 10 data asli dengan benar).
- Kerentanan Kritis & Sangat Tinggi: *Recall* masing-masing bergeser menjadi 44,00% dan 33,33%.

Penerapan teknik SMOTE yang diisolasi khusus pada data pelatihan terbukti mampu memberikan keseimbangan performa model Decision Tree dengan nilai akurasi keseluruhan sebesar 56,00%.

Berdasarkan hasil pengujian tersebut, analisis terhadap metrik presisi (*class precision*) menunjukkan bahwa model memiliki tingkat ketepatan prediksi paling tinggi pada kelas minoritas Kerentanan Rendah, yaitu mencapai 80,00% (4 dari 5 data yang diprediksi sebagai Kerentanan Rendah adalah tepat). Sementara itu, tingkat presisi untuk kategori Kerentanan Sedang/Tinggi, Kerentanan Sangat Tinggi, dan Kerentanan Kritis masing-masing mencatatkan nilai sebesar 56,45%, 54,55%, dan 50,00%.

Dilihat dari sisi daya deteksi atau sensitivitas (*class recall*), penyeimbangan data latih menggunakan SMOTE berhasil mempertahankan kestabilan model dalam mengidentifikasi sampel-sampel aktual pada kelas minoritas. Model berhasil meningkatkan *recall* Kerentanan Rendah menjadi 40,00% dengan mengklasifikasikan 4 dari 10 data asli secara benar, serta mencatatkan nilai *recall* sebesar 44,00% (11 dari 25 data) untuk Kerentanan Kritis dan 33,33% (6 dari 18 data) untuk Kerentanan Sangat Tinggi. Adapun daya deteksi tertinggi tetap berada pada kelas mayoritas Kerentanan Sedang/Tinggi dengan nilai *recall* sebesar 74,47% (35 dari 47 data).

Sebagai metrik harmonik yang menggambarkan keseimbangan antara presisi dan *recall*, nilai F1-Score untuk masing-masing kelas terdistribusi secara proposional. Kelas Kerentanan Sedang/Tinggi memperoleh nilai F1-Score tertinggi sebesar 64,22%, diikuti oleh kelas minoritas Kerentanan Rendah sebesar 53,33%, Kerentanan Kritis sebesar 46,81%, dan Kerentanan Sangat Tinggi sebesar 41,38%. Secara keseluruhan, model Decision Tree yang dilatih dengan penyeimbangan SMOTE pada *train set* dan diuji pada *test set* murni ini menghasilkan rata-rata makro (*macro average*) F1-Score sebesar 51,44%. Hal ini mengonfirmasi bahwa penyeimbangan data latih berhasil menekan bias model terhadap kelas mayoritas tanpa memicu manipulasi data (*data leakage*) pada tahap evaluasi.

Pembahasan dan Implikasi Temuan

Diskusi Teoritis dan Dinamika Faktor Prediktor

Hasil pemodelan integrasi K-Means dan *Decision Tree* menunjukkan bahwa variabel Persepsi Mengenai Kejahatan Siber menjadi *root node* atau faktor pemicu utama dalam mengklasifikasikan tingkat kerentanan pengguna internet. Temuan ini sejalan dengan kerangka kerja *Protection Motivation Theory* (PMT) dalam studi keamanan siber terkini, yang menegaskan bahwa persepsi individu terhadap ancaman (*perceived severity*) dan tingkat risiko (*perceived vulnerability*) secara langsung mengarahkan tindakan proteksi diri mereka (Demetz *et al*., 2023). Namun, tingginya tingkat persepsi ancaman tidak selalu berkorelasi positif dengan perilaku keamanan yang ideal.

Sebagian responden yang merasa sangat paham dan setuju mengenai risiko kejahatan siber justru berada pada kategori Kerentanan Kritis hingga Sangat Tinggi. Fenomena ini mengonfirmasi keberadaan *Optimism Bias* dan *Overconfidence Effect* pada pengguna internet modern, di mana individu yang merasa memiliki pengetahuan teknis cenderung menyepelekan prosedur keamanan dasar karena merasa dirinya bukan target utama serangan (Zwilling *et al*., 2022).

Fenomena Kerentanan Berdasarkan Demografi dan Pengetahuan

Penggabungan faktor usia, jenis kelamin, dan tingkat pengetahuan memperlihatkan pola kerentanan yang kompleks pada pengguna internet. Pada kelompok usia produktif (20–35 tahun), responden perempuan dengan pengetahuan "Sangat Paham" teridentifikasi berada dalam kategori Kerentanan Sangat Tinggi, sementara responden laki-laki berada dalam kategori Kerentanan Kritis. Temuan ini memperkuat bukti adanya *knowledge-behavior gap* (kesenjangan antara pengetahuan dan perilaku) pada ekosistem digital saat ini, di mana tingginya literasi teori tidak otomatis terkonversi menjadi disiplin perilaku keamanan (Sarker, 2023). Tingginya mobilitas dan frekuensi aktivitas digital pada kelompok usia produktif secara tidak langsung memperluas area permukaan serangan (*attack surface*) melalui transaksi daring dan interaksi media sosial yang intensif (Alqahtani & Watson, 2024). Sebaliknya, responden dengan tingkat persepsi "Cukup Setuju" dan rentang usia muda (<20 tahun)

cenderung berada pada kelas Kerentanan Rendah, yang menunjukkan bahwa keterbatasan akses terhadap transaksi finansial dan sistem sensitif membatasi tingkat keterpaparan risiko siber mereka.

Komparasi dengan Studi Terdahulu dan Evaluasi Pemodelan

Temuan dalam penelitian ini mengonfirmasi efektivitas integrasi teknik klasterisasi dan klasifikasi dalam memetakan domain keamanan siber. Penelitian oleh Han *et al*. (2022) membuktikan bahwa penggabungan K-Means dan *Decision Tree* sangat andal dalam mengekstraksi pola dari data heterogen multidimensi. Meski demikian, ketidakseimbangan kelas (*class imbalance*) sering kali menjadi kendala utama yang memicu timbulnya *bias* prediksi pada kelas mayoritas (Krawczyk & Woźniak, 2021). Penerapan teknik SMOTE pada penelitian ini terbukti efektif mengatasi kendala tersebut, di mana nilai *recall* untuk kelas minoritas (Kerentanan Rendah) mengalami peningkatan pesat dari 30.00% menjadi 63.83%. Walaupun akurasi secara keseluruhan berada pada angka 59.12%, peningkatan sensitivitas deteksi pada kelas minoritas merupakan aspek paling krusial dalam analisis keamanan siber. Dalam mitigasi risiko siber, kesalahan gagal mendeteksi entitas rentan (kesalahan Tipe II) memiliki dampak kerugian yang jauh lebih fatal dibandingkan kesalahan klasifikasi minoritas (Fernández-Navarro *et al*., 2023b).

Implikasi Praktis dan Kebijakan Literasi Siber

Secara praktis, hasil penelitian ini memberikan implikasi penting bagi formulasi kebijakan literasi digital dan mitigasi risiko siber. Temuan penelitian menegaskan bahwa program edukasi keamanan siber tidak boleh lagi berfokus hanya pada transfer pengetahuan teoretis (*knowledge-based approach*), melainkan harus bertransformasi menjadi pendekatan intervensi perilaku berbasis segmentasi (*behavioral-based intervention*). Pemerintah dan penyedia layanan digital disarankan untuk merancang kampanye kesadaran siber (*cybersecurity awareness*) yang disesuaikan dengan profil demografi dan tingkat kerentanan spesifik. Sebagai contoh, edukasi untuk kelompok usia produktif perlu difokuskan pada kewaspadaan rekayasa sosial (*social engineering*) serta teknik *phishing*, sedangkan untuk kelompok pengguna muda fokus edukasi dapat diarahkan pada perlindungan data pribadi dan kebersihan privasi digital.

KESIMPULAN

Berdasarkan integrasi K-Means Clustering dan Decision Tree C4.5, penelitian ini berhasil mengidentifikasi profil serta karakteristik pembeda kerentanan pengguna internet terhadap kejahatan siber. K-Means mengelompokkan 100 responden ke dalam empat kategori, yaitu Kerentanan Moderat sebesar 47%, Kerentanan Rendah sebesar 10%, Kerentanan Sangat Tinggi atau *Behavioral Gap* sebesar 18%, dan Kerentanan Kritis sebesar 25%. Hasil ini menunjukkan bahwa sebagian besar responden berada pada kategori rentan hingga sangat rentan, yang berkaitan dengan belum konsistennya penerapan perilaku proteksi digital. Model Decision Tree mengidentifikasi Persepsi Mengenai Kejahatan Siber sebagai atribut pembeda utama, diikuti oleh usia, jenis kelamin, dan pengetahuan mengenai kejahatan siber. Pola tersebut menunjukkan adanya indikasi *knowledge-behavior gap*, khususnya pada kelompok usia produktif, ketika pengetahuan keamanan yang tinggi belum selalu diikuti perilaku digital yang aman. Penerapan SMOTE yang hanya dilakukan pada data latih menghasilkan akurasi model sebesar 56,00% dan *macro F1-score* sebesar 51,44%, serta meningkatkan *recall* kelas Kerentanan Rendah dari 30,00% menjadi 40,00% dengan presisi 80,00%. Dengan performa tersebut, aturan keputusan yang dihasilkan perlu dipahami sebagai pola eksploratif awal. Temuan penelitian memberikan dasar bagi pengembangan program literasi keamanan siber yang tidak hanya berorientasi pada peningkatan pengetahuan, tetapi juga pada perubahan perilaku berdasarkan karakteristik kelompok pengguna. Penelitian selanjutnya disarankan memperluas ukuran dan cakupan sampel, mengevaluasi algoritma klasifikasi lain seperti Random Forest dan XGBoost, serta menambahkan atribut prediktor yang lebih spesifik untuk meningkatkan kemampuan model dalam memetakan kerentanan pengguna internet.

REFERENSI

Alqahtani, H., & Watson, T. (2024). Human factors in cybersecurity: Evaluating attack surfaces in modern digital environments. *Journal of Cybersecurity and Privacy*, 4(1), 45-62.

Almufareh, M. F., Al-Ghamdi, A. S., & Mohd, F. (2024). Comprehensive evaluation of clustering validity indices for optimal cluster selection in high-dimensional security data. *Journal of Big Data*, 11(1), Article 45. https://doi.org/10.1186/s40537-024-00912-x

Amoo, O. O., Adeyemo, A. B., & Omidiora, E. O. (2022). Cybercrime risk assessment using hybrid clustering and machine learning techniques. *Computers & Security*, 118, Article 102731. https://doi.org/10.1016/j.cose.2022.102731

Ariyaningsih, S., Andrianto, A. A., Kusuma, A. surya, & Rezi. (2023). Korelasi kejahatan siber dengan percepatan digitalisasi di Indonesia. *Justisia: Jurnal Ilmu Hukum*, *1*. https://doi.org/10.56457/jjih.v1i1.38

Daeng, Y., Levin, J., Razzaq Prayudha, M., Putri Ramadhani, N., Noverto, Imanuel, S., & Virgio. (2023). Analisis penerapan sistem keamanan siber terhadap kejahatan siber di Indonesia. *INNOVATIVE: Journal of Social Science Research*, *3*.

Demetz, L., Taylor, S., & Stevens, M. (2023). Rethinking protection motivation theory in digital safety: A contemporary review. *Computers & Security*, 128, 103150.

Fernández-Navarro, F., Hervás-Martínez, C., & Gutierrez, P. A. (2023). Advanced classification models and resampling techniques for imbalanced data in risk assessment. *IEEE Transactions on Cybernetics*, 53(4), 2410–2422. https://doi.org/10.1109/TCYB.2022.3189001

Fernández-Navarro, F., Hervás-Martínez, C., & Gutierrez, P. A. (2023). Cost-sensitive learning and class imbalance approaches for cybersecurity risk detection. *Information Fusion*, 91, 210-224. https://doi.org/10.1016/j.inffus.2023.101804

Han, J., Kamber, M., & Pei, J. (2022). *Data mining: concepts and techniques* (4th ed.). Morgan Kaufmann. https://doi.org/10.1016/C2013-0-18660-6

ITU. (2026). *Hampir tiga perempat populasi terhubung ke internet.* ITU. https://www.itu.int/itu-d/reports/statistics/2025/10/15/ff25-internet-use/

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. STHDA.

Krawczyk, B., & Woźniak, M. (2021). Recent advances in imbalanced data classification: Algorithms, applications, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 2980-2995.

Muningsih, E. (2022). Kombinasi metode K-Means dan Decision Tree dengan perbandingan kriteria dan split data. *Jurnal Teknoinfo*, 16(1), 112–119. https://doi.org/10.33365/jti.v16i1.1520

Prusty, S. R., Patnaik, S., & Dash, S. K. (2022). Integration of unsupervised clustering and decision tree algorithms for behavioral risk stratification in cybersecurity. *Decision Support Systems*, 154, Article 113702. https://doi.org/10.1016/j.dss.2021.113702

Sari, S. N., Annisa, P., Rahma, A. N. D., Ritonga, R. P., & Utomo, D. P. (2025). Teknik data mining dengan menggunakan algoritma Decision Tree untuk mengetahui pola pemahaman mahasiswa pada matakuliah pemrograman. *Bulletin of Information Technology (BIT)*, 6(4), 301–309. https://doi.org/10.47065/bit.v6i4.1210

Sarker, I. H. (2023). Cybersecurity data science: Best practices, algorithms, and real-world applications. *Journal of Network and Computer Applications*, 211, 103560.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, 8, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Syafrizal, M., Ashari, M. W., & Cane, M. A. M. (2025). Kerangka Penelitian: Menuju Penyelesaian Studi Dan Penelitian Keamanan Siber. *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, *10*(2), 1446–1462. https://doi.org/10.36341/rabit.v10i2.6715

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means clustering method and Elbow method for identification of the best cluster number. *IOP Conference Series: Materials Science and Engineering*, 336(1), Article 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Zwilling, M., Kietzmann, J., & Techatassanasoontorn, A. A. (2022). The overconfidence phenomenon in cybersecurity behaviors: An empirical investigation. *Computers in Human Behavior*, 134, 107312.