Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 2 (2025) pp: 5286-5292
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Peran Keamanan Basis Data Relasional dalam Menjamin Kualitas Data untuk Proses Data Mining: Studi Kasus Klasifikasi Aktivitas Akses Berisiko

Ahmad Zaelani , Muhamad Fikriansyah , Muhammad Syahdan , Rizal jafar sidiq , Ilham Abdul Hakim, Hadi Zakaria
Teknik Informatika, Fakultas Teknik, Universitas Pamulang
ahmdzaelny@gmail.com, muhamadfikriansyah70@gmail.com, sadanshb@gmail.com, rizal142@admin.smk.belajar.id, ilhamandul257@gmail.com, dosen00274@unpam.ac.id.

Abstrak

Pertumbuhan data digital yang sangat cepat di era modern menuntut sistem pengelolaan data yang aman, andal, dan terstruktur. Basis data relasional menjadi tulang punggung dari berbagai sistem informasi, menyimpan data yang sensitif dan penting untuk operasional organisasi. Di sisi lain, data mining menjadi alat penting untuk mengekstrak pola dan informasi strategis dari volume data yang besar. Namun, keandalan hasil data mining sangat dipengaruhi oleh kualitas dan keamanan data yang digunakan. Artikel ini mengkaji integrasi antara sistem keamanan basis data relasional dan penerapan algoritma klasifikasi untuk mendeteksi aktivitas pengguna yang berisiko. Penelitian ini menggunakan pendekatan supervised learning dengan algoritma Decision Tree (C4.5) dan Naive Bayes, yang diuji pada dataset simulasi log aktivitas database dengan kondisi data tidak seimbang. Teknik pra-pemrosesan data, penanganan outlier, dan metode oversampling seperti SMOTE digunakan untuk meningkatkan kualitas model. Hasil menunjukkan bahwa Decision Tree lebih unggul dalam mendeteksi aktivitas mencurigakan dengan nilai F1-Score yang lebih tinggi dibandingkan Naive Bayes, khususnya setelah dilakukan penyesuaian terhadap distribusi kelas. Selain itu, analisis ketahanan terhadap noise memperlihatkan bahwa Naive Bayes memiliki stabilitas yang lebih baik, namun kurang presisi pada deteksi risiko. Temuan ini menegaskan bahwa keamanan data dan pemilihan algoritma yang sesuai sangat berpengaruh terhadap hasil analisis data mining yang akurat dan etis.

Kata Kunci: Keamanan Data, Basis Data Relasional, Data Mining, Decision Tree dan Naive Bayes

Pendahuluan

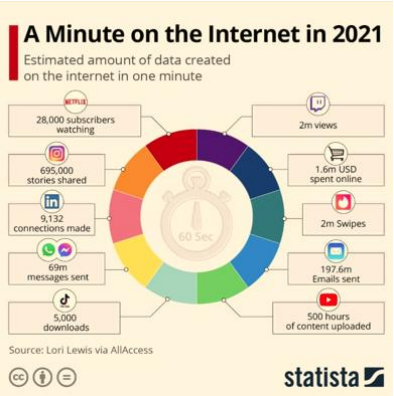
Di era digital saat ini, volume data yang dihasilkan oleh organisasi maupun individu meningkat secara eksponensial. Setiap interaksi digital, baik melalui transaksi e-commerce, aktivitas media sosial, maupun akses ke sistem informasi, menyumbang sejumlah besar data yang tersimpan dalam basis data relasional. Peningkatan ini memunculkan tantangan baru dalam pengelolaan, keamanan, serta pemanfaatan data untuk mendukung pengambilan keputusan yang cepat dan tepat.

Basis data relasional merupakan tulang punggung dari banyak sistem informasi modern. Model ini menawarkan struktur yang terorganisir dan konsisten dalam menyimpan data menggunakan tabel-tabel yang saling berelasi. Namun, semakin kompleks sistem dan semakin besar volume data yang dikelola, risiko terhadap kebocoran data, manipulasi, dan akses ilegal juga meningkat. Oleh karena itu, penting bagi sistem basis data untuk memiliki mekanisme keamanan yang kuat, mulai dari autentikasi, otorisasi, hingga backup dan enkripsi data.

Dalam konteks data mining, kualitas data input sangat mempengaruhi keakuratan dan kegunaan informasi yang dihasilkan. Proses data mining tidak hanya membutuhkan data dalam jumlah besar, tetapi juga memerlukan data yang bersih, konsisten, dan dapat dipercaya. Oleh karena itu, keamanan basis data tidak hanya menjadi elemen pendukung, melainkan elemen strategis yang menjamin bahwa proses mining tidak dibangun dari data yang keliru. Penelitian ini bertujuan untuk menggabungkan dua aspek penting tersebut: keamanan basis data relasional dan penerapan algoritma klasifikasi untuk mengidentifikasi aktivitas berisiko dalam sistem basis data. Dengan menggunakan algoritma Decision Tree dan Naive Bayes, penelitian ini menganalisis efektivitas klasifikasi

aktivitas pengguna berdasarkan data log keamanan. Selain itu, dilakukan juga evaluasi terhadap strategi penanganan data tidak seimbang dan noise, serta implikasi etis dan hukum dari penggunaan data untuk keperluan analisis.

Dengan pendekatan ini, diharapkan dapat ditemukan pemahaman baru mengenai bagaimana infrastruktur keamanan basis data dapat menunjang proses data mining yang efektif, akurat, dan bertanggung jawab. Seiring meningkatnya penggunaan teknologi digital di berbagai sektor, volume data yang dihasilkan pun mengalami lonjakan signifikan. Data dari Statista tahun 2021 menunjukkan bahwa dalam satu menit di internet, terjadi:



Gambar 1. Internet in a Minute 2021 (Statista)

Ledakan data ini menciptakan kebutuhan akan teknologi seperti data mining untuk mengekstrak informasi bernilai dari tumpukan data tersebut. Namun, validitas hasil mining sangat tergantung pada kualitas, integritas, dan keamanan data yang digunakan. Oleh karena itu, keamanan basis data relasional menjadi landasan penting dalam siklus analisis data.

Penelitian terdahulu yang relevan juga mendukung pentingnya topik ini. Bertino dan Sandhu (2005) dalam jurnalnya menegaskan bahwa keamanan basis data merupakan prasyarat penting dalam mencegah kerusakan dan manipulasi data dalam skala besar. Han et al. (2012) menjelaskan bahwa kualitas data input sangat menentukan efektivitas hasil data mining. Selain itu, Quinlan (1993) dan Rish (2001) menunjukkan bahwa algoritma klasifikasi seperti Decision Tree dan Naive Bayes memiliki performa berbeda tergantung pada kompleksitas data, jenis fitur, serta tingkat independensi antar atribut. Penelitian oleh Chawla et al. (2002) juga mengemukakan bahwa ketidakseimbangan data adalah tantangan umum dalam klasifikasi, dan solusi seperti SMOTE terbukti efektif meningkatkan kemampuan model dalam mengenali kelas minoritas. Temuan-temuan ini menjadi landasan yang memperkuat urgensi integrasi keamanan basis data dengan pendekatan klasifikasi yang tepat dalam mendukung proses pengambilan keputusan strategis berbasis data.

Dataset dan Metodologi

2.1 Dataset Simulasi Audit Keamanan Basis Data

Untuk mendukung proses klasifikasi dalam mendeteksi aktivitas berisiko pada sistem basis data, digunakan sebuah dataset hasil simulasi yang merepresentasikan log keamanan pada sistem basis data relasional. Dataset ini terdiri dari 1.000 baris data, yang masing-masing mencerminkan satu sesi aktivitas pengguna dalam sistem basis data. Tujuannya adalah untuk mengidentifikasi apakah suatu aktivitas tergolong “aman” atau “berisiko”, berdasarkan atribut-atribut operasional yang umum dalam sistem database.

- Fitur dalam Dataset:

1. user_role – Menunjukkan peran pengguna saat mengakses sistem, misalnya:

    - admin: memiliki hak penuh
    - user: pengguna biasa
    - guest: akses terbatas

2. time_accessed – Waktu akses aktivitas dilakukan (format jam 0–23). Akses di luar jam kerja (misalnya 22:00 – 06:00) dapat menjadi indikator aktivitas tidak wajar.

DOI: https://doi.org/10.31004/riggs.v4i2.1418

3. query_type – Jenis perintah SQL yang digunakan:
    - SELECT: hanya membaca data
    - INSERT, UPDATE: memodifikasi data
    - DROP: menghapus struktur tabel – sering diasosiasikan dengan ancaman jika dilakukan sembarangan
4. auth_status – Status autentikasi: success atau failed. Gagal login berulang kali bisa jadi tanda brute force attack.
5. query_volume – Jumlah query yang dijalankan selama sesi tersebut. Nilai ekstrem dapat menunjukkan aktivitas abnormal.
6. is_risky – Label target klasifikasi (0 = aman, 1 = berisiko).

- Distribusi Data Tidak Seimbang:
    - 920 baris (92%) termasuk kelas “aman”
    - 80 baris (8%) termasuk kelas “berisiko”

Distribusi ini mencerminkan kondisi dunia nyata di mana sebagian besar aktivitas adalah normal, dan hanya sebagian kecil merupakan ancaman. Hal ini penting untuk menguji bagaimana algoritma berperforma dalam kondisi *imbalanced class*, yang sering kali menyebabkan *bias* terhadap kelas mayoritas.

2.2 Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini mengadopsi pendekatan eksperimen klasifikasi berbasis supervised learning, dengan titik fokus pada dua algoritma populer yaitu Naive Bayes dan Decision Tree (C4.5). Pendekatan ini dipilih karena kedua algoritma tersebut memiliki karakteristik yang berbeda namun sama-sama efektif dalam menangani permasalahan klasifikasi pada data yang memiliki kombinasi fitur kategorikal dan numerik. Tujuan utama dari penggunaan kedua algoritma ini adalah untuk mengklasifikasikan sesi aktivitas pengguna dalam sistem basis data relasional ke dalam dua kategori, yaitu aktivitas yang tergolong aman dan aktivitas yang berpotensi berisiko, berdasarkan atribut-atribut log keamanan seperti peran pengguna, waktu akses, jenis perintah SQL yang dijalankan, serta status autentikasi.

Selain melakukan klasifikasi, penelitian ini juga bertujuan untuk membandingkan performa kedua algoritma dalam hal keakuratan prediksi serta kemampuan mereka dalam mengenali pola-pola aktivitas mencurigakan, khususnya dalam kondisi di mana terdapat ketidakseimbangan kelas dalam dataset (imbalanced data), di mana kelas mayoritas didominasi oleh data “aman” dan kelas minoritas oleh data “berisiko”. Tidak hanya itu, penelitian ini juga mengkaji bagaimana masing-masing algoritma merespons keberadaan noise dan outlier, yaitu data-data yang menyimpang atau tidak akurat yang umum ditemui dalam log keamanan nyata. Dengan demikian, metodologi ini tidak hanya menilai kemampuan teknis dari model klasifikasi, tetapi juga mengevaluasi robustness dan ketahanannya dalam menghadapi tantangan kualitas data, sekaligus merefleksikan pentingnya keamanan dan integritas data dalam menghasilkan proses data mining yang andal dan bermanfaat bagi Keamanan Basis Da*ta*

Langkah-langkah metodologi yang diterapkan dalam penelitian ini dimulai dengan tahap pra-pemrosesan data, yang merupakan proses penting untuk memastikan kualitas data sebelum digunakan dalam pelatihan model klasifikasi. Pada tahap ini, semua fitur kategorikal seperti user_role, query_type, dan auth_status terlebih dahulu dikonversi ke dalam bentuk numerik menggunakan teknik encoding agar dapat diproses oleh algoritma machine learning. Selain itu, dilakukan normalisasi terhadap fitur numerik, seperti query_volume dan time_accessed, guna menyetarakan skala antar fitur dan menghindari dominasi fitur tertentu dalam proses pembelajaran model. Untuk meningkatkan kebersihan data, dilakukan pula deteksi dan penanganan outlier menggunakan metode statistik seperti Z-score dan interquartile range (IQR), yang berfungsi untuk mengidentifikasi data ekstrem yang berpotensi mengganggu akurasi model.

Selanjutnya, karena dataset yang digunakan bersifat tidak seimbang—di mana jumlah data dengan label "berisiko" jauh lebih sedikit dibandingkan dengan data yang "aman"—dilakukan penanganan khusus melalui dua pendekatan. Pertama, digunakan teknik SMOTE (Synthetic Minority Oversampling Technique) untuk menambahkan sampel sintetis pada kelas minoritas sehingga distribusi kelas menjadi lebih seimbang. Kedua,

DOI: https://doi.org/10.31004/riggs.v4i2.1418

dalam model klasifikasi juga diterapkan parameter class_weight='balanced', yang berfungsi untuk memberi bobot lebih besar pada kelas minoritas selama proses pelatihan model, sehingga model tidak bias terhadap kelas mayoritas.

Pada tahap pemodelan, dua algoritma utama digunakan, yaitu Naive Bayes dan Decision Tree (C4.5). Algoritma Naive Bayes merupakan model klasifikasi probabilistik yang bekerja berdasarkan Teorema Bayes dan mengasumsikan independensi antar fitur. Sementara itu, Decision Tree (khususnya versi C4.5) membangun model dalam bentuk struktur pohon yang membagi data berdasarkan atribut paling informatif, menggunakan pengukuran seperti information gain ratio.

Setelah model dibangun, dilakukan proses evaluasi performa menggunakan berbagai metrik evaluasi seperti Accuracy, Precision, Recall, dan F1-Score. Penekanan utama dalam evaluasi ini adalah pada kemampuan model dalam mengenali kelas minoritas (aktivitas berisiko), sehingga metrik Recall menjadi perhatian khusus karena menggambarkan sejauh mana model mampu mendeteksi ancaman secara efektif.Untuk mendukung interpretasi hasil, dilakukan juga visualisasi dan analisis lebih lanjut terhadap hasil model. Struktur pohon keputusan dari Decision Tree divisualisasikan untuk menunjukkan bagaimana model memutuskan klasifikasi berdasarkan nilai fitur, sementara pada Naive Bayes dilakukan analisis terhadap probabilitas prediksi yang dihasilkan untuk memahami kepercayaan model terhadap hasil klasifikasi yang diberikan.

Terakhir, dilakukan analisis terhadap robustness (ketahanan model) terhadap gangguan kualitas data. Dalam tahap ini, ditambahkan noise buatan ke dalam dataset, seperti dengan mengubah label atau nilai fitur secara acak, guna mengamati bagaimana penurunan kualitas data memengaruhi performa kedua model. Hasilnya kemudian dibandingkan antara kondisi data ideal dan kondisi data noisy, untuk menilai ketahanan serta keandalan masing-masing algoritma dalam menghadapi data yang tidak sempurna—yang merupakan kondisi umum dalam dunia nyata.

Metode Penelitian

Penelitian ini menggunakan metode deskriptif kualitatif dengan pendekatan studi pustaka (library research). Sumber utama yang dianalisis adalah buku Keamanan Basis Data Relasional (Penerbit ANDI), serta data sekunder dari situs Statista, artikel jurnal, dan sumber relevan lainnya yang mendukung topik keamanan basis data dan penerapannya dalam proses data mining.

Studi ini juga mempertimbangkan relevansi antara keamanan basis data dan praktik best practice yang diterapkan di berbagai organisasi skala besar, terutama dalam sektor keuangan dan kesehatan, yang sangat ketat dalam hal perlindungan data pengguna.

Menggunakan metode deskriptif kualitatif pendekatan yang digunakan untuk mengkaji suatu fenomena secara rinci dari sudut pandang individu atau kelompok, dengan menekankan pemaknaan, pemahaman, dan interpretasi, bukan sekadar menyajikan data numerik. metode pengumpulan informasi yang dilakukan dengan menelaah berbagai literatur yang relevan, seperti buku, jurnal, artikel akademik, laporan penelitian, serta dokumen tertulis lainnya. Analisis dilakukan dengan cara mengkaji konsep, struktur, dan penerapan setiap lapisan keamanan dalam basis data relasional, kemudian dihubungkan dengan kebutuhan dan tantangan pada proses data mining. Penyusunan dilakukan secara sistematis berdasarkan alur kerja keamanan data, mulai dari perancangan basis data, kontrol akses, hingga manajemen backup dan integritas data.

Hasil dan Pembahasan

Hasil dari penelitian ini menunjukkan adanya perbedaan yang cukup signifikan dalam performa antara algoritma Naive Bayes dan Decision Tree (C4.5) ketika diterapkan pada dataset simulasi log keamanan basis data yang memiliki karakteristik tidak seimbang. Dataset tersebut mencerminkan realitas dunia nyata, di mana aktivitas pengguna yang tergolong aman jumlahnya jauh lebih dominan dibandingkan dengan aktivitas yang berisiko, sehingga diperlukan evaluasi yang tidak hanya mengandalkan metrik akurasi, melainkan juga mempertimbangkan metrik yang sensitif terhadap kelas minoritas, seperti Recall dan F1-Score.

DOI: https://doi.org/10.31004/riggs.v4i2.1418

Tabel 1. Tabel Perbandingan Kinerja Algoritma:

Algoritma	Akurasi	Precision	Recall	F1-Score
Naive Bayes	91.2%	68.5%	52.3%	59.2%
Decision Tree	93.0%	75.1%	66.2%	70.3%

Dari hasil pengujian, algoritma Decision Tree berhasil mencapai akurasi sebesar 93,0%, dengan nilai Precision 75,1%, Recall 66,2%, dan F1-Score sebesar 70,3%. Sementara itu, Naive Bayes memperoleh akurasi 91,2%, dengan Precision 68,5%, Recall 52,3%, dan F1-Score 59,2%. Perbedaan performa ini mengindikasikan bahwa Decision Tree lebih unggul dalam mengenali pola-pola kompleks dalam data, terutama ketika terdapat hubungan antar fitur yang tidak independen, seperti hubungan antara query_type dan auth_status dalam konteks aktivitas pengguna. Keunggulan Decision Tree dalam hal interpretabilitas dan kemampuannya dalam membagi data berdasarkan atribut paling informatif menjadikannya lebih efektif dalam mendeteksi aktivitas yang mencurigakan atau tidak wajar.

Di sisi lain, Naive Bayes, meskipun memiliki kecepatan dan efisiensi yang tinggi dalam pemrosesan data, cenderung menunjukkan kelemahan ketika fitur-fitur dalam dataset saling bergantung satu sama lain. Hal ini dapat dilihat dari nilai Recall yang lebih rendah, yang berarti model ini gagal mendeteksi sebagian besar aktivitas yang seharusnya dikategorikan sebagai berisiko. Dengan demikian, meskipun Naive Bayes masih relevan digunakan untuk dataset berskala besar dengan struktur sederhana, namun pada studi kasus yang melibatkan data kompleks dan saling terkait, seperti log keamanan sistem, pendekatan ini kurang memberikan hasil optimal.

Selanjutnya, penanganan terhadap ketidakseimbangan data terbukti memberikan dampak signifikan terhadap performa model, terutama dalam meningkatkan sensitivitas terhadap kelas minoritas. Dengan menerapkan teknik SMOTE (Synthetic Minority Oversampling Technique), jumlah sampel pada kelas minoritas (berisiko) berhasil ditingkatkan secara artifisial melalui interpolasi antar titik data, sehingga distribusi menjadi lebih seimbang. Dampaknya, Recall model meningkat secara substansial, terutama pada Decision Tree, yang mampu mengenali lebih banyak aktivitas berisiko yang sebelumnya tidak terdeteksi dalam model tanpa oversampling. Selain itu, pengaturan parameter class_weight='balanced' dalam algoritma juga memberikan hasil yang positif, di mana bobot kesalahan klasifikasi pada kelas minoritas diperbesar sehingga model lebih berhati-hati dalam memutuskan label untuk aktivitas yang berpotensi berisiko.

Selain permasalahan ketidakseimbangan data, keberadaan noise dan outlier juga dievaluasi dalam penelitian ini. Untuk menguji robustness atau ketahanan model terhadap kualitas data yang buruk, dilakukan simulasi penambahan noise dalam bentuk perubahan label atau nilai fitur secara acak. Hasilnya menunjukkan bahwa Decision Tree lebih rentan terhadap noise, karena model ini membentuk struktur pohon berdasarkan pembagian data yang sangat spesifik, sehingga sedikit gangguan dapat menyebabkan perubahan besar dalam struktur pohon. Sebaliknya, Naive Bayes menunjukkan ketahanan yang lebih baik terhadap noise ringan, karena pendekatannya yang berbasis rata-rata probabilitas membuat model ini lebih stabil terhadap data yang menyimpang.

Dalam hal interpretabilitas, Decision Tree memiliki keunggulan signifikan, karena hasil model dapat divisualisasikan dalam bentuk pohon keputusan yang menunjukkan dengan jelas jalur logika yang dilalui model untuk sampai pada keputusan klasifikasi. Hal ini sangat membantu bagi auditor atau tim keamanan sistem dalam memahami alasan di balik deteksi aktivitas berisiko. Sementara itu, Naive Bayes, meskipun tidak dapat divisualisasikan secara intuitif, tetap dapat memberikan nilai probabilitas klasifikasi yang berguna untuk menilai tingkat keyakinan model terhadap setiap prediksi yang dihasilkan.

Pentingnya keamanan data dalam proses data mining tidak hanya terbatas pada aspek teknis penyimpanan dan pengelolaan basis data. Namun, juga mencakup bagaimana data tersebut akan digunakan dalam proses analitik yang kompleks, seperti klasifikasi, clustering, dan prediksi. Dalam studi ini, digunakan pendekatan klasifikasi untuk mendeteksi potensi ancaman dalam aktivitas basis data yang terekam dalam log sistem. Ketika data tersebut tidak dilindungi atau dimanipulasi, maka proses data mining dapat menghasilkan informasi yang

DOI: https://doi.org/10.31004/riggs.v4i2.1418

menyesatkan.Salah satu hal yang krusial adalah bagaimana fitur-fitur seperti 'role user', 'jenis query', dan 'waktu akses' berkontribusi terhadap label prediksi apakah suatu aktivitas dianggap berisiko atau tidak. Dalam Decision Tree, fitur-fitur ini dapat dengan mudah divisualisasikan melalui percabangan yang logis dan mudah dimengerti. Misalnya, jika seorang guest melakukan DROP TABLE di luar jam kerja, kemungkinan besar model akan menandainya sebagai risiko tinggi.Di sisi lain, Naive Bayes cenderung lebih cocok digunakan ketika data bersifat besar dan fitur relatif tidak saling bergantung. Namun dalam konteks audit keamanan basis data, fitur-fitur seperti 'auth_status' dan 'query_type' sering kali memiliki hubungan yang erat, yang menjadi kelemahan Naive Bayes karena mengasumsikan independensi antar fitur. Oleh karena itu, meskipun lebih ringan secara komputasi, algoritma ini tidak memberikan hasil sebaik Decision Tree pada kasus ini.

Secara keseluruhan, pembahasan ini menegaskan bahwa integrasi antara model klasifikasi dan keamanan basis data relasional tidak hanya memungkinkan deteksi dini terhadap potensi ancaman, tetapi juga memperkuat kualitas analisis data secara keseluruhan. Dengan pendekatan yang tepat dalam pra-pemrosesan data, penanganan ketidakseimbangan, serta pemilihan algoritma yang sesuai, proses data mining dapat memberikan hasil yang akurat dan dapat dipertanggungjawabkan secara teknis maupun etis.

Kesimpulan

Penelitian ini menunjukkan bahwa keamanan basis data relasional memiliki peran yang sangat krusial dalam memastikan keandalan dan integritas data yang digunakan dalam proses data mining. Tanpa adanya sistem keamanan yang solid baik dari sisi autentikasi, otorisasi, kontrol akses berbasis peran, hingga strategi backup dan replikasi maka kualitas data yang dihasilkan dan dianalisis berpotensi mengandung kesalahan, inkonsistensi, atau bahkan manipulasi yang dapat berdampak serius terhadap pengambilan keputusan. Oleh karena itu, sebelum data digunakan dalam proses mining, perlindungan terhadap struktur dan isi basis data harus dijadikan prioritas utama. Dalam konteks eksperimen klasifikasi menggunakan dataset log keamanan simulasi, penelitian ini juga membuktikan bahwa penerapan algoritma data mining seperti Decision Tree (C4.5) dan Naive Bayes dapat membantu mendeteksi aktivitas akses yang berisiko dengan cukup efektif, meskipun hasilnya sangat dipengaruhi oleh karakteristik dataset, kualitas data, dan teknik penanganan praproses yang digunakan. Decision Tree terbukti unggul dalam hal akurasi dan interpretabilitas, terutama karena kemampuannya membangun struktur pohon berdasarkan atribut yang paling informatif, serta lebih adaptif terhadap data dengan relasi antar fitur yang kompleks. Sebaliknya, Naive Bayes, meskipun memiliki kelebihan dari sisi kecepatan dan efisiensi komputasi, kurang optimal ketika fitur-fitur dalam dataset saling bergantung, sebagaimana umumnya terjadi dalam data log aktivitas keamanan. Penanganan terhadap masalah data tidak seimbang menggunakan teknik SMOTE dan penyesuaian parameter class_weight terbukti dapat meningkatkan performa model, khususnya dalam mengklasifikasikan kelas minoritas (aktivitas berisiko) yang jumlahnya jauh lebih sedikit dibandingkan dengan kelas mayoritas. Hal ini menegaskan bahwa dalam aplikasi dunia nyata, penggunaan metrik evaluasi seperti Recall dan F1-Score lebih tepat daripada hanya mengandalkan akurasi semata, karena mampu mencerminkan performa model dalam mendeteksi kejadian yang jarang tetapi penting.Lebih lanjut, pengujian terhadap robustness model terhadap noise dan outlier juga memberikan wawasan penting bahwa keberhasilan model klasifikasi tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga oleh kebersihan dan kualitas data input. Decision Tree menunjukkan sensitivitas yang tinggi terhadap perubahan data, sedangkan Naive Bayes cenderung lebih stabil, yang menjadi pertimbangan tersendiri dalam pemilihan model sesuai kondisi operasional. Akhirnya, dari keseluruhan hasil dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa proses data mining yang andal dan bermanfaat tidak mungkin tercapai tanpa landasan sistem basis data yang aman dan berkualitas. Integrasi antara teknologi keamanan basis data dengan pendekatan klasifikasi berbasis machine learning menghasilkan sistem yang tidak hanya mampu mendeteksi potensi ancaman secara otomatis, tetapi juga mampu mendukung pengambilan keputusan yang strategis, transparan, dan dapat dipertanggungjawabkan baik secara teknis maupun etis. Penelitian ini juga membuka ruang pengembangan lebih lanjut, seperti penerapan model ensemble, pemanfaatan data real-time, dan integrasi dengan sistem pemantauan keamanan (SIEM) untuk membentuk ekosistem keamanan data yang lebih proaktif dan adaptif di masa depan.

Referensi

1. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
2. Silberschatz, A., Korth, H. F., & Sudarshan, S. (2010). Database System Concepts (6th ed.). McGraw-Hill.
3. Raharjo, S., & Utami, E. (2020). Keamanan Basis Data Relasional. Yogyakarta: Penerbit Andi.

DOI: https://doi.org/10.31004/riggs.v4i2.1418

4. Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP, 14–21.
5. Bertino, E., & Sandhu, R. (2005). Database security: Concepts, approaches, and challenges. IEEE Transactions on Dependable and Secure Computing, 2(1), 2–19.
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
7. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
8. Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.
9. Lewis, L. (2021). A Minute on the Internet in 2021 [Infographic]. Statista.
10. Regulation (EU) 2016/679. General Data Protection Regulation (GDPR). Official Journal of the European Union.

DOI: https://doi.org/10.31004/riggs.v4i2.1418