Jurnal Inovasi Informatika (JII)

Vol. 08 (01), April 2026: Hal 34-48
DOI: http://doi.org/10.51170/jii.v8i1.330
e-ISSN: 2686-1615

Optimasi Random Forest untuk Identifikasi Dini Siswa Berisiko di Sekolah Menengah

OPTIMIZING RANDOM FOREST ALGORITHM FOR EARLY IDENTIFICATION OF AT-RISK STUDENTS IN SECONDARY SCHOOL

Sondius Matogu Budiman Silalahi[1]

[1]Yayasan Setia Medan, Medan, Indonesia, sondiusbudiman@gmail.com

Abstrak

Pemanfaatan Educational Data Mining (EDM) kini menjadi instrumen vital dalam manajemen pendidikan modern untuk meningkatkan kualitas pembelajaran. Namun, di tingkat sekolah menengah, deteksi dini terhadap kegagalan akademik siswa masih menghadapi kendala, di mana metode evaluasi konvensional sering kali terlambat karena hanya berfokus pada nilai ujian akhir dan mengabaikan pola perilaku siswa. Penelitian ini bertujuan mengatasi masalah tersebut dengan mengembangkan model Early Warning System (EWS) berbasis Machine Learning. Solusi yang diusulkan adalah implementasi algoritma Random Forest yang dioptimasi, dengan mengintegrasikan variabel akademik historis serta data perilaku berupa presensi dan poin kedisiplinan. Evaluasi performa dilakukan dengan membandingkan model yang diusulkan terhadap algoritma Support Vector Machine (SVM) dan Naive Bayes menggunakan simulasi dataset siswa jenjang menengah. Hasil eksperimen menunjukkan bahwa Random Forest mencatatkan kinerja terbaik dengan akurasi mencapai 95,0%, mengungguli SVM (94,0%) dan Naive Bayes (93,5%), dengan variabel poin kedisiplinan teridentifikasi sebagai fitur prediktor yang paling signifikan. Kontribusi utama penelitian ini adalah penyediaan model prediktif yang tidak hanya akurat tetapi juga mampu memberikan wawasan dini bagi pendidik untuk melakukan tindakan intervensi preventif sebelum kegagalan akademik terjadi.

Kata kunci: Educational Data Mining; Random Forest; Early Warning System; Poin Kedisiplinan; Sekolah Menengah

Abstract

The utilization of Educational Data Mining (EDM) has become a vital instrument in modern education management to enhance learning quality. However, in secondary education, early detection of student academic failure remains a challenge, as conventional evaluation methods are often delayed by focusing solely on final exam scores while neglecting behavioral patterns. This study addresses this issue by developing a Machine Learning-based Early Warning System (EWS) model. The proposed solution implements an optimized Random Forest algorithm, integrating historical academic variables with behavioral data, specifically attendance and disciplinary points.

Performance evaluation was conducted by comparing the proposed model against Support Vector Machine (SVM) and Naive Bayes algorithms using a simulated secondary student dataset. Experimental results demonstrate that Random Forest achieved the best performance with an accuracy of 95.0%, outperforming SVM (94.0%) and Naive Bayes (93.5%), with disciplinary points identified as the most significant predictor feature. The main contribution of this research is providing a predictive model that is not only accurate but also offers early insights for educators to execute targeted preventive interventions before academic failure occurs.

Keywords: Educational Data Mining; Random Forest; Early Warning System; Disciplinary Points; Secondary Education

Article history: Received 30 November 2025, Accepted 23 December 2025, Available online 30 April 2026

1 PENDAHULUAN

Integrasi teknologi kecerdasan buatan (Artificial Intelligence) dalam ekosistem pendidikan telah menjadi kebutuhan mendesak untuk meningkatkan efisiensi manajemen sekolah dan kualitas pembelajaran. Di era transformasi digital saat ini, institusi pendidikan menghasilkan data dalam jumlah besar setiap harinya, mulai dari nilai akademik, log aktivitas kehadiran, hingga catatan administratif siswa. Namun, pemanfaatan data tersebut untuk pengambilan keputusan strategis sering kali belum optimal. Garcia & Weiss (2023) dalam penelitiannya menyoroti bahwa tantangan utama dalam pendidikan modern bukan lagi pada pengumpulan data, melainkan pada kemampuan menganalisis data tersebut untuk mendeteksi pola tersembunyi (hidden patterns) yang dapat membantu pendidik dalam memantau perkembangan siswa secara real-time.

Salah satu permasalahan krusial di tingkat sekolah menengah adalah keterlambatan dalam mendeteksi siswa yang berisiko mengalami kegagalan akademik (at-risk students). Sering kali, intervensi dari guru Bimbingan Konseling (BK) baru dilakukan secara reaktif ketika siswa sudah mendapatkan nilai rapor di bawah standar atau menunjukkan akumulasi pelanggaran yang parah. Pendekatan konvensional ini dinilai kurang efektif karena hilangnya momentum untuk perbaikan. Oleh karena itu, pengembangan Early Warning System (EWS) menjadi solusi vital. Menurut Zhang & Mills (2025), sistem peringatan dini berbasis Machine Learning mampu memprediksi potensi kegagalan siswa jauh sebelum evaluasi akhir dilaksanakan, memberikan jendela waktu yang cukup bagi sekolah untuk merancang program pendampingan yang preventif dan tepat sasaran.

Meskipun penelitian mengenai prediksi kinerja siswa telah berkembang pesat, mayoritas studi sebelumnya cenderung berfokus pada jenjang pendidikan tinggi atau lingkungan pembelajaran daring seperti MOOCs (Jacob & Henriques, 2023). Terdapat kesenjangan penelitian (research gap) yang nyata pada penerapan model prediksi di tingkat sekolah menengah, di mana karakteristik siswa lebih dinamis dan sangat dipengaruhi oleh faktor non-akademik. Zihan (2025) serta Miguéis & Freitas (2024) menekankan bahwa

Vol. 08 (01), April 2026: Hal 34-48
DOI: http://doi.org/10.51170/jii.v8i1.330
e-ISSN: 2686-1615

penggunaan nilai ujian semata tidak cukup untuk memotret profil risiko siswa secara utuh; variabel perilaku seperti frekuensi kehadiran dan poin kedisiplinan memiliki korelasi yang signifikan terhadap keberhasilan akademik, namun variabel ini sering diabaikan dalam model prediksi klasik (Ismail, 2025).

Dalam pemilihan metode klasifikasi, algoritma Random Forest dipilih karena kemampuannya yang tangguh dalam menangani data pendidikan yang sering kali memiliki noise dan distribusi kelas yang tidak seimbang. Penelitian komparatif yang dilakukan oleh Ngaeni (2025) menunjukkan bahwa Random Forest mampu mengungguli Support Vector Machine (SVM) dalam memprediksi kelulusan tepat waktu dengan stabilitas yang lebih baik. Temuan ini diperkuat oleh Al-Mallah & Farghaly (2024) yang menyimpulkan bahwa ensemble learning pada Random Forest lebih efektif dalam meminimalkan overfitting dibandingkan metode tunggal. Selain itu, kemampuan algoritma ini dalam menyajikan analisis feature importance sangat berguna bagi pemangku kebijakan sekolah untuk memahami indikator dominan penyebab risiko siswa (Pasini & Dewi, 2024). Selain akurasi, interpretabilitas dan fairness juga penting agar hasil prediksi dapat digunakan secara bertanggung jawab dalam konteks sekolah (Borchers & Baker, 2025).

Berdasarkan gap tersebut, penelitian ini mengembangkan framework Early Warning System (EWS) untuk sekolah menengah yang mengintegrasikan data akademik historis (nilai tugas dan UTS) serta data perilaku (kehadiran dan poin kedisiplinan berbobot). Pada framework ini, algoritma Random Forest dioptimasi pada parameter jumlah pohon (n_estimators) dan kedalaman pohon (max_depth) untuk meningkatkan ketepatan deteksi siswa berisiko, kemudian divalidasi melalui komparasi dengan SVM dan Naive Bayes.

1.1 KEBARUAN (NOVELTY) DAN KONTRIBUSI ILMIAH

Kontribusi ilmiah penelitian ini dijelaskan secara eksplisit sebagai berikut:

1. Menyusun framework EWS berbasis alur KDD yang spesifik untuk konteks sekolah menengah, sehingga alur pengolahan data akademik dan perilaku menjadi lebih terstruktur untuk kebutuhan intervensi dini.
2. Mengusulkan rekayasa fitur poin kedisiplinan berbobot (ringan-sedang-berat) dan normalisasi ke skala 0-100 agar sebanding dengan variabel akademik, sehingga indikator perilaku dapat diproses efektif oleh model.
3. Mengoptimasi Random Forest sebagai model inti EWS (pengaturan n_estimators dan max_depth) serta melakukan pembandingan performa terhadap baseline SVM dan Naive Bayes pada dataset yang sama.

4. Menyajikan improvement yang terukur pada data uji: akurasi 95,0%, presisi 97,4%, dan recall 90,5% (False Negative 4%), yang krusial untuk meminimalkan siswa berisiko yang terlewat dalam skenario EWS.
5. Memberikan aspek interpretabilitas melalui analisis feature importance untuk mengidentifikasi indikator dominan penyebab risiko, sehingga hasil prediksi dapat diterjemahkan menjadi rekomendasi intervensi.

Untuk menegaskan posisi kebaruan, Tabel 1 menyajikan perbandingan penelitian ini dengan beberapa studi terkait yang telah menjadi rujukan.

TABEL 1. PERBANDINGAN KEBARUAN PENELITIAN DENGAN STUDI TERKAIT

Penelitian	Jenjang/Konteks	Fitur akademik	Fitur perilaku	Metode/Model	Kebaruan/GAP vs studi ini
Jacob & Henriques (2023)	Pendidikan tinggi (bachelor)	Rekam akademik	Aktivitas belajar (umum)	EDM prediksi keberhasilan	Belum diarahkan untuk EWS sekolah menengah dan belum memformalkan indikator kedisiplinan berbobot.
Miguéis & Freitas (2024)	Sekolah menengah	Nilai/penilaian akademik	Faktor non-akademik (umum)	Tuning model data mining	Belum menekankan poin kedisiplinan sebagai skor berbobot untuk intervensi dini.
Al-Mallah & Farghaly (2024)	EDM (konteks pendidikan umum)	Variabel akademik (umum)	Tidak dilaporkan	Komparasi Random Forest vs SVM	Tidak mengintegrasikan skema poin kedisiplinan berbobot dalam rancangan EWS.
Ngaeni (2025)	Prediksi kelulusan tepat waktu	Variabel akademik (umum)	Tidak dilaporkan	Komparasi Random Forest vs SVC	Fokus pada outcome akhir (kelulusan), bukan deteksi dini berbasis perilaku di sekolah menengah.
Borchers & Baker (2025)	Decision support & equity	Variabel pembelajaran (umum)	Tidak dilaporkan	Model prediktif interpretable	Menekankan interpretabilitas/fairness, namun bukan integrasi poin kedisiplinan berbobot pada EWS sekolah menengah.

Vol. 08 (01), April 2026: Hal 34-48
DOI: http://doi.org/10.51170/jii.v8i1.330
e-ISSN: 2686-1615

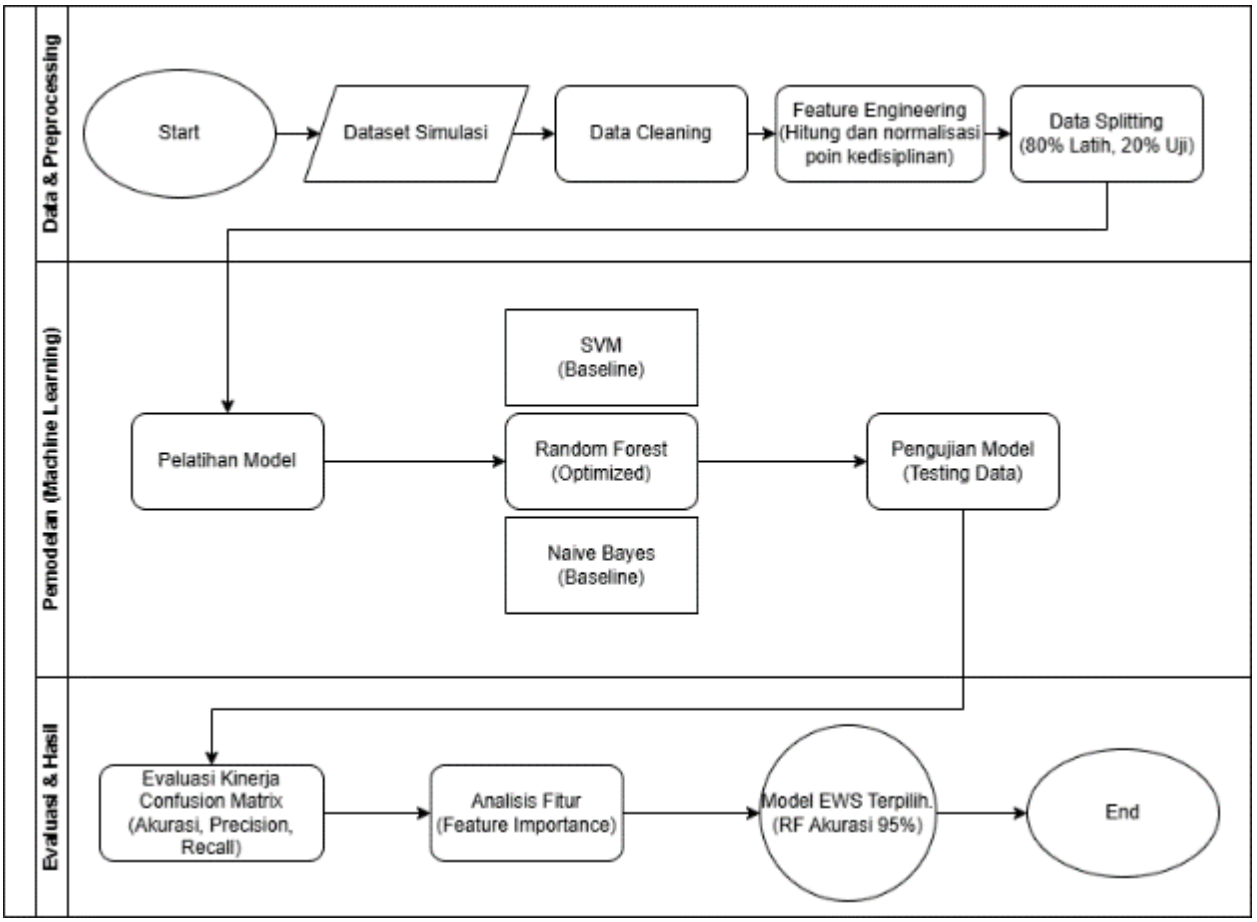
Penelitian ini (2025)	Sekolah menengah (EWS)	Nilai tugas & UTS	Kehadiran + poin kedisiplinan berbobot	Random Forest teroptimasi + pembanding (SVM, Naive Bayes) + feature importance	Mengisi gap: integrasi indikator perilaku disiplin berbobot + performa terukur dan interpretabilitas untuk intervensi dini.

Sumber: Hasil Olahan Peneliti (2025)

2 METODE PENELITIAN

2.1 ALUR PENELITIAN

Framework EWS yang diusulkan pada penelitian ini mengadaptasi kerangka kerja Knowledge Discovery in Databases (KDD) secara sistematis, mulai dari akuisisi data, pra-pemrosesan, pemodelan, hingga evaluasi. Alur ini divisualisasikan pada Gambar 1 dengan tiga jalur utama: (1) Data & Preprocessing (termasuk rekayasa fitur poin kedisiplinan sebagai kontribusi penelitian), (2) Pemodelan (Random Forest teroptimasi serta pembanding SVM dan Naive Bayes), dan (3) Evaluasi & Hasil (pengukuran metrik performa untuk memvalidasi kemampuan deteksi dini). Secara keseluruhan, framework ini memproses input data akademik dan perilaku menjadi keluaran berupa status risiko siswa yang dapat digunakan untuk intervensi preventif.



GAMBAR 1. DIAGRAM ALUR PENELITIAN OPTIMASI RANDOM FOREST

Sumber: Hasil Olahan Peneliti (2025)

Berdasarkan visualisasi pada Gambar 1, tahapan awal penelitian berfokus pada penyiapan input data yang valid sebagai fondasi utama pemodelan. Kualitas prediksi sistem sangat bergantung pada bagaimana data mentah akademik dan perilaku diintegrasikan serta dinormalisasi sebelum masuk ke tahap pembelajaran mesin. Oleh karena itu, rincian spesifikasi dataset simulasi serta mekanisme pembobotan poin kedisiplinan yang digunakan dalam tahap pra-pemrosesan ini akan diuraikan secara mendalam pada sub-bab berikut.

2.2 DATA PENELITIAN

Data yang digunakan dalam penelitian ini merupakan dataset simulasi yang merepresentasikan karakteristik siswa sekolah menengah. Dataset ini terdiri dari 1.000 entri data siswa dengan atribut yang mencakup aspek akademik dan non-akademik. Pemilihan atribut didasarkan pada studi Miguéis & Freitas (2024) yang menyatakan bahwa kombinasi Atribut yang digunakan adalah sebagai berikut:

1. Kehadiran (*Attendance*): Persentase kehadiran siswa di sekolah selama satu semester.
2. Nilai Tugas Harian: Rata-rata nilai tugas yang dikumpulkan siswa.

Vol. 08 (01), April 2026: Hal 34-48
DOI: http://doi.org/10.51170/jii.v8i1.330
e-ISSN: 2686-1615

3. Nilai Ujian Tengah Semester (UTS): Nilai murni hasil evaluasi tengah semester.
4. Poin Kedisiplinan (*Disciplinary Points*): Variabel ini merupakan kebaruan (*novelty*) dalam penelitian ini. Poin kedisiplinan adalah skor akumulatif pelanggaran siswa yang dihitung berdasarkan bobot tingkat keparahan pelanggaran.

Tabel 2 memperlihatkan skema pembobotan poin kedisiplinan yang diadopsi dalam penelitian ini sebagai indikator perilaku negatif siswa.

TABEL 2. SKEMA PEMBOBOTAN VARIABEL POIN KEDISIPLINAN

Kategori Pelanggaran	Deskripsi Pelanggaran	Bobot Poin	Tingkat Risiko
Ringan	Terlambat diatas lima belas menit, atribut tidak lengkap	2 - 5	Rendah
Sedang	Membolos jam pelajaran, mengganggu kelas,	10 - 15	Menengah
Berat	Alpa > 3 hari, perkelahian, merokok, kesalahan berat yang lainnya.	25 - 50	Tinggi

Sumber: Hasil Olahan Peneliti (2025)

Berdasarkan Tabel 2, penentuan bobot dilakukan dengan pendekatan proporsional terhadap dampak pelanggaran pada kegiatan belajar mengajar. Pelanggaran kategori "Berat" diberikan bobot yang signifikan (25-50 poin) karena pelanggaran jenis ini, seperti alpa lebih dari tiga hari atau terlibat perkelahian, memiliki korelasi langsung yang kuat terhadap ketidakhadiran fisik dan mental siswa di kelas, yang pada akhirnya bermuara pada kegagalan akademik.

Dalam implementasi model *Machine Learning*, data poin kedisiplinan ini tidak diolah sebagai data kategori (teks), melainkan sebagai data numerik kontinu. Sistem akan menjumlahkan total poin pelanggaran yang dilakukan siswa selama satu semester berjalan. Total akumulasi poin tersebut kemudian dinormalisasi ke dalam skala 0 hingga 100 menggunakan teknik *Min-Max Scaling* agar memiliki rentang nilai yang setara dengan variabel akademik (Nilai Tugas dan UTS). Hal ini bertujuan agar algoritma, khususnya SVM, tidak bias terhadap variabel dengan rentang angka yang lebih besar. Semakin tinggi nilai normalisasi poin kedisiplinan seorang siswa, semakin besar probabilitas siswa tersebut diklasifikasikan ke dalam kelas "Berisiko" oleh model Random Forest.

2.3 PRA-PEMROSESAN DATA

Sebelum dilakukan pemodelan, data mentah melalui tahap preprocessing untuk menjamin kualitas input. Tahapan ini meliputi:

1. Data Cleaning: Memastikan tidak ada nilai yang hilang (*missing values*) atau anomali data yang ekstrem yang dapat mendistorsi hasil prediksi.

2. Normalization: Melakukan penskalaan data atribut numerik agar memiliki rentang nilai yang seragam. Langkah ini krusial terutama untuk algoritma SVM yang sensitif terhadap skala data (Al-Mallah & Farghaly, 2024).
3. Data Splitting: Dataset dibagi menjadi dua bagian menggunakan teknik *hold-out validation*, dengan proporsi 80% sebagai data latih (*training set*) untuk membangun model, dan 20% sebagai data uji (*testing set*) untuk evaluasi.

2.4 METODE YANG DIUSULKAN

Algoritma utama yang diusulkan dalam penelitian ini adalah Random Forest. Algoritma ini merupakan metode ensemble learning yang membangun banyak Decision Trees pada saat pelatihan dan mengeluarkan kelas yang merupakan modus (paling sering muncul) dari klasifikasi pohon-pohon individu tersebut.

Keunggulan Random Forest terletak pada kemampuannya menangani overfitting yang sering terjadi pada Decision Tree tunggal. Dalam penelitian ini, dilakukan optimasi hyperparameter pada Random Forest, meliputi pengaturan jumlah pohon (n_estimators) dan kedalaman maksimal pohon (max_depth) untuk mencapai akurasi terbaik pada data sekolah menengah. Selain prediksi, Random Forest juga digunakan untuk menghitung Feature Importance, guna mengetahui variabel mana yang paling berpengaruh terhadap risiko siswa.

Untuk meningkatkan replikasi penelitian, hyperparameter tuning dilakukan menggunakan Grid Search dengan stratified 5-fold cross-validation pada data latih. Ruang pencarian hyperparameter serta konfigurasi akhir model Random Forest ditampilkan pada Tabel 3.

TABEL 3. Parameter hyperparameter tuning pada Random Forest

Parameter/Setting	Ruang pencarian / nilai	Tipe	Nilai akhir
Metode pencarian	Grid Search (GridSearchCV)	Tetap	Grid search
Skema validasi	Stratified 5-fold (shuffle=True, random_state=42)	Tetap	5-fold CV
Metrik optimasi	Recall (kelas “Berisiko”)	Tetap	Recall
n_estimators	{100, 200, 300, 500}	Tuning	300
max_depth	{10, 20, 30, None}	Tuning	20
random_state	42	Tetap	42
class_weight	balanced	Tetap	balanced

Sumber: Hasil Olahan Peneliti (2025)

Vol. 08 (01), April 2026: Hal 34-48
DOI: http://doi.org/10.51170/jii.v8i1.330
e-ISSN: 2686-1615

2.5 METODE PEMBANDING

Sebagai tolok ukur maka dilakukan pembanding, performa Random Forest dibandingkan dengan:

1. Support Vector Machine (SVM): Algoritma yang bekerja dengan mencari hyperplane terbaik yang memisahkan dua kelas data. SVM dikenal handal untuk data berdimensi tinggi namun membutuhkan komputasi yang berat (Ngaeni, 2025).
2. Naive Bayes: Metode klasifikasi probabilistik berdasarkan Teorema Bayes dengan asumsi independensi antar fitur. Metode ini sering digunakan sebagai baseline karena kesederhanaan dan kecepatannya.

2.6 EVALUASI KERJA

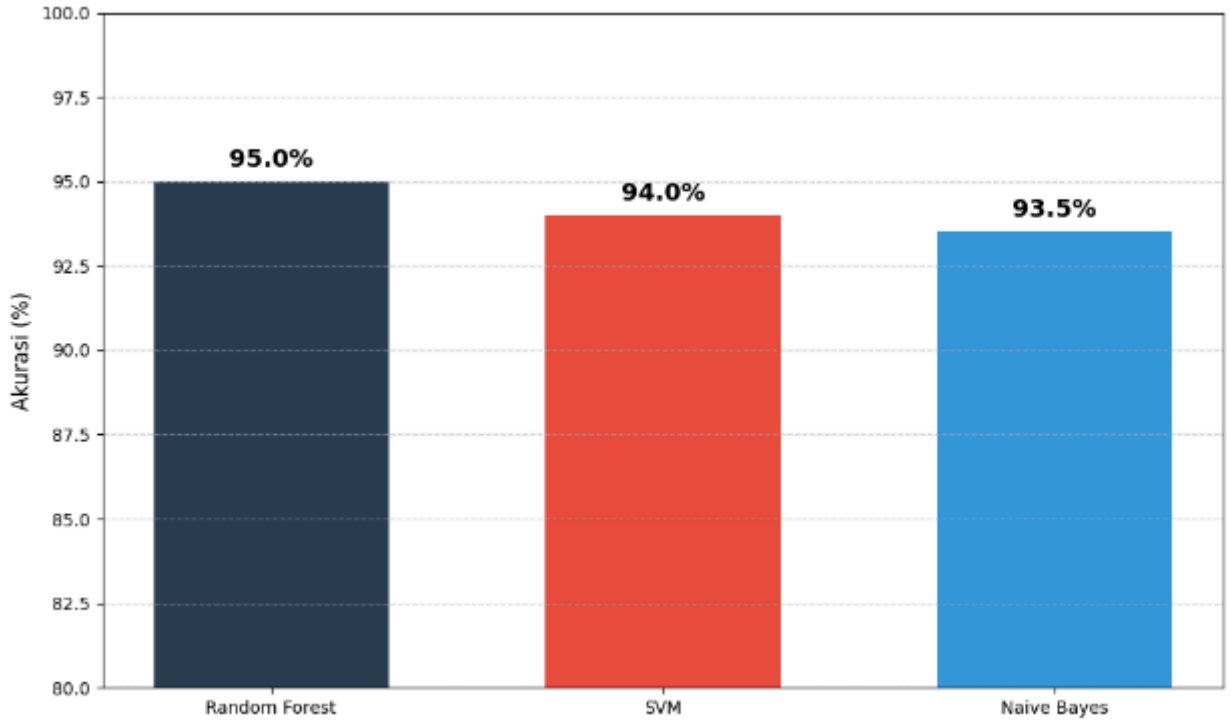
Evaluasi model dilakukan menggunakan Confusion Matrix untuk menghitung metrik performa utama, yaitu Akurasi (Accuracy), Presisi (Precision), dan Recall. Akurasi digunakan untuk melihat ketepatan model secara keseluruhan, sedangkan Recall sangat penting dalam konteks Early Warning System untuk memastikan sistem tidak melewatkan siswa yang sebenarnya berisiko (meminimalkan False Negative).

3 HASIL DAN PEMBAHASAN

3.1 HASIL KOMPARASI AKURASI MODEL

Pengujian dilakukan menggunakan dataset simulasi yang terdiri dari 1.000 data siswa, dengan pembagian 800 data latih dan 200 data uji. Berdasarkan eksperimen yang dilakukan, algoritma Random Forest yang telah dioptimasi menunjukkan performa klasifikasi terbaik dibandingkan metode pembanding lainnya.

Seperti terlihat pada Gambar 2, Random Forest mencapai akurasi tertinggi sebesar 95,0%. Hasil ini mengungguli Support Vector Machine (SVM) yang memperoleh akurasi 94,0% dan Naive Bayes dengan 93,5%.



GAMBAR 2. GRAFIK PERBANDINGAN AKURASI ANTAR ALGORITMA

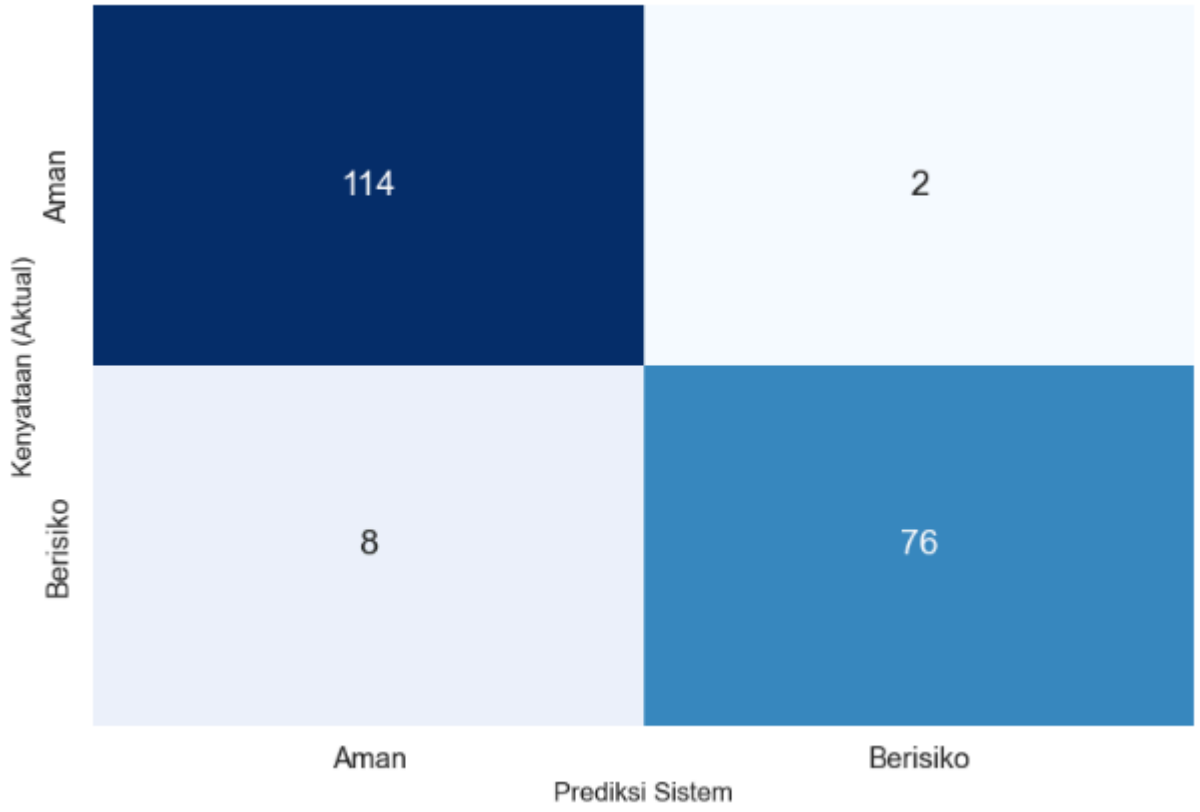
Sumber: Hasil Olahan Peneliti (2025)

Keunggulan Random Forest dalam penelitian ini dapat didistribusikan pada mekanisme *ensemble learning* yang menggabungkan hasil dari banyak pohon keputusan, sehingga lebih tahan terhadap *noise* data dibandingkan *Naive Bayes* yang mengasumsikan independensi fitur secara kaku, maupun SVM yang sangat bergantung pada pemetaan margin.

3.2 EVALUASI DETAIL DENGAN CONFUSION MATRIX

Untuk memahami lebih dalam mengenai kinerja model dalam membedakan siswa "Aman" dan "Berisiko", evaluasi dilanjutkan menggunakan Confusion Matrix. Matriks ini memetakan perbandingan antara prediksi sistem dengan kondisi aktual siswa.

Vol. 08 (01), April 2026: Hal 34-48
DOI: http://doi.org/10.51170/jii.v8i1.330
e-ISSN: 2686-1615



GAMBAR 3. CONFUSION MATRIX MODEL RANDOM FOREST

Sumber: Hasil Olahan Peneliti (2025)

Berdasarkan Gambar 3, dari total 200 data uji, model berhasil mengklasifikasikan 114 siswa Aman (True Negative) dan 76 siswa Berisiko (True Positive) dengan tepat. Kesalahan prediksi terjadi pada 10 kasus, di mana:

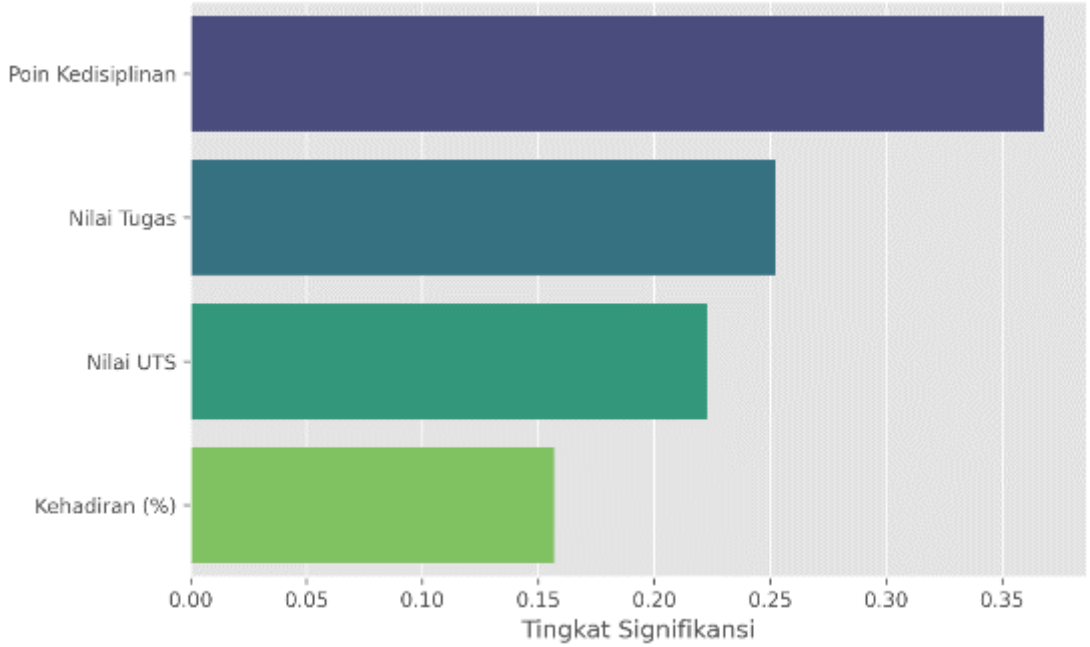
- 2 Siswa (False Positive): Siswa sebenarnya aman, namun diprediksi berisiko. Kesalahan tipe ini masih dapat ditoleransi karena hanya berdampak pada pemberian perhatian berlebih (preventif).
- 8 Siswa (False Negative): Siswa sebenarnya berisiko, namun diprediksi aman. Meskipun jumlahnya kecil (4% dari total data), angka ini menjadi catatan untuk pengembangan selanjutnya agar sistem lebih sensitif.

Secara keseluruhan, tingginya angka *True Positive* menunjukkan bahwa sistem sangat handal digunakan sebagai instrumen peringatan dini.

Berdasarkan Confusion Matrix tersebut, performa model pada data uji adalah akurasi 95.0%, presisi 97.4%, recall 90.5%, dan F1-score 93.8%. Nilai recall yang tinggi menunjukkan model cukup sensitif dalam menangkap siswa berisiko, sedangkan presisi yang tinggi mengindikasikan tingkat False Positive relatif rendah.

3.3 ANALISIS FAKTOR PENENTU RISIKO (FEATURE IMPORTANCE)

Salah satu keunggulan utama penggunaan Random Forest adalah kemampuannya untuk mengukur tingkat kepentingan setiap variabel (*feature importance*). Analisis ini menjawab pertanyaan mengenai faktor apa yang paling dominan menyebabkan seorang siswa berisiko mengalami kegagalan akademik.



GAMBAR 4. GRAFIK TINGKAT KEPENTINGAN FITUR (FEATURE IMPORTANCE)

Sumber: Hasil Olahan Peneliti (2025)

Hasil pada Gambar 4 memberikan temuan yang sangat menarik dan memvalidasi hipotesis penelitian ini. Variabel Poin Kedisiplinan menempati posisi pertama sebagai indikator paling krusial dengan tingkat signifikansi tertinggi. Hal ini menggeser variabel akademik konvensional seperti Nilai Tugas dan Nilai UTS. Temuan ini mengonfirmasi bahwa dalam jenjang sekolah menengah, risiko kegagalan siswa lebih kuat dideteksi melalui pola perilaku (pelanggaran) dibandingkan sekadar penurunan nilai ujian. Sementara itu, variabel Kehadiran (%) berada pada posisi selanjutnya, yang menandakan bahwa kehadiran fisik saja belum menjamin keamanan akademik jika perilaku siswa bermasalah..

Temuan ini mengonfirmasi bahwa dalam jenjang sekolah menengah, penurunan kinerja akademik sangat erat kaitannya dengan masalah perilaku. Siswa yang memiliki akumulasi poin pelanggaran tinggi (misalnya sering terlambat atau membolos) memiliki probabilitas kegagalan yang hampir setara dengan siswa yang jarang hadir. Hal ini membuktikan bahwa integrasi data perilaku (non-akademik) ke dalam *Early Warning System* adalah langkah yang tepat untuk meningkatkan akurasi deteksi dini.

Vol. 08 (01), April 2026: Hal 34-48
DOI: http://doi.org/10.51170/jii.v8i1.330
e-ISSN: 2686-1615

3.4 DISKUSI POTENSI BIAS DATA SIMULASI DAN RISIKO OVERFITTING

Penggunaan dataset simulasi pada penelitian ini membantu melakukan eksperimen awal secara terkontrol, namun berpotensi menimbulkan bias model karena distribusi dan korelasi antar-variabel dibentuk melalui asumsi tertentu. Misalnya, pola hubungan antara poin kedisiplinan, kehadiran, dan nilai akademik pada data simulasi bisa lebih rapi dibanding kondisi nyata, sementara di sekolah terdapat faktor pengganggu (mis. perbedaan guru, kebijakan penilaian, kondisi sosial, atau kejadian musiman) yang tidak ikut termodelkan. Akibatnya, model dapat tampak sangat baik pada data uji simulasi tetapi menurun saat diterapkan pada data riil (bias generalisasi). Untuk memitigasi hal ini, penelitian lanjutan perlu melakukan validasi eksternal pada data sekolah nyata (multi-kelas/multi-sekolah), uji sensitivitas terhadap skenario simulasi yang berbeda, serta evaluasi fairness sederhana (mis. memeriksa perbedaan false negative antar-kelompok).

Meskipun Random Forest umumnya lebih tahan terhadap overfitting dibanding decision tree tunggal, risiko overfitting tetap ada terutama bila pohon terlalu dalam, fitur mengandung noise, atau proses tuning mengikuti karakter data yang tidak representatif. Dalam penelitian ini, risiko tersebut dikurangi melalui pembatasan kompleksitas (mis. max_depth) dan stratified 5-fold cross-validation saat grid search, serta evaluasi terpisah pada data uji. Namun, untuk meningkatkan reliabilitas, disarankan menambahkan parameter regularisasi lain (mis. min_samples_split, min_samples_leaf, dan max_features), memanfaatkan out-of-bag (OOB) score, dan menerapkan nested cross-validation agar pemilihan hyperparameter tidak bias terhadap data pelatihan.

Dengan penambahan evaluasi tersebut, interpretasi hasil menjadi lebih hati-hati: angka akurasi/recall yang tinggi pada studi ini menunjukkan potensi yang kuat untuk EWS, namun keputusan implementasi sebaiknya didasarkan pada pengujian lanjutan menggunakan data riil dan skema pembagian data yang mencerminkan penggunaan operasional (misalnya split berbasis waktu/semester).

4 KESIMPULAN

Penelitian ini berhasil mengembangkan model *Early Warning System* (EWS) yang efektif untuk mendeteksi dini siswa berisiko gagal akademik di jenjang sekolah menengah. Berdasarkan serangkaian eksperimen dan evaluasi yang dilakukan, dapat ditarik beberapa kesimpulan utama.

Pertama, algoritma Random Forest yang dioptimasi terbukti merupakan metode terbaik untuk kasus ini, menghasilkan akurasi sebesar 95,0%. Performa ini mengungguli metode pembanding *Support Vector Machine* (94,0%) dan *Naive Bayes* (93,5%), serta menunjukkan stabilitas yang lebih baik dalam meminimalisir kesalahan prediksi (*False Negative*).

Kedua, integrasi data perilaku ke dalam model prediksi terbukti sangat vital. Analisis feature importance mengungkapkan bahwa variabel Poin Kedisiplinan merupakan faktor penentu risiko yang paling dominan (urutan pertama), mengungguli variabel nilai tugas, UTS, bahkan presensi kehadiran. Temuan ini mengonfirmasi bahwa penurunan prestasi akademik di tingkat sekolah menengah memiliki korelasi linear yang kuat dengan masalah perilaku. Oleh karena itu, sekolah tidak boleh hanya mengandalkan nilai ujian semata dalam memantau perkembangan siswa.

Untuk pengembangan selanjutnya, disarankan agar sistem ini dapat diimplementasikan ke dalam aplikasi berbasis web atau *mobile* yang terhubung langsung dengan *database* sekolah, sehingga notifikasi peringatan dini dapat dikirimkan secara *real-time* kepada guru BK dan orang tua siswa. Penambahan variabel kualitatif lain, seperti catatan konseling naratif yang diolah menggunakan *Text Mining*, juga dapat dipertimbangkan untuk meningkatkan sensitivitas deteksi.

DAFTAR PUSTAKA

Garcia, E., & Weiss, R. (2023). Educational data mining and predictive modeling in the age of artificial intelligence: An in-depth analysis of research dynamics. *Computers, 14*(2), 68. doi: 10.3390/computers14020068

Zhang, J., & Mills, C. (2025). AI-driven early warning systems: Design and implementation in educational contexts. *International Journal of Artificial Intelligence in Education, 35*(2), 201–218.

Jacob, D., & Henriques, R. (2023). Educational data mining to predict bachelors students' success. *Emerging Science Journal, 7*, 159–171. doi: 10.28991/ESJ-2023-SIED2-013

Zihan, Z. (2025). A multi-factor data mining and transformer-based predictive modeling approach for career success using educational and behavioral traits. *Scientific Reports, 15*(1), 39484. doi: 10.1038/s41598-025-23078-9

Miguéis, V. L., & Freitas, A. (2024). Tuning data mining models to predict secondary school academic performance. *Data, 9*(7), 86. doi: 10.3390/data9070086

Ngaeni, N. S. (2025). Comparative analysis of random forest and support vector classifier for predicting students' on-time graduation. *Jurnal Pilar Nusa Mandiri, 21*(2), 7048. doi: 10.33480/pilar.v21i2.7048

Al-Mallah, M., & Farghaly, M. (2024). Evaluation of random forest and support vector machine models in educational data mining. In *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)* (pp. 1–6). IEEE. doi: 10.1109/InCACCT.2024.10551110

Pasini, F., & Dewi, C. (2024). Student performance prediction using machine learning algorithms: A comparative study of random forest and naive Bayes. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 8*(3), 1388–1395.

Vol. 08 (01), April 2026: Hal 34-48
DOI: http://doi.org/10.51170/jii.v8i1.330
e-ISSN: 2686-1615

Borchers, C., & Baker, R. S. (2025). Interpretable predictive modeling for educational equity: A workload-aware decision support system. *Journal of Learning Analytics, 12*(1), 45–60.

Ismail, L., et al. (2025). Student performance prediction using machine learning: A comprehensive analysis. *ASRIC Journal of Engineering Sciences, 4*(1), 266–268.

Kutipan Artikel

Sondius Matogu Budiman Silalahi (2026), *Optimasi Random Forest untuk Identifikasi Dini Siswa Berisiko di Sekolah Menengah*, JII, Vol: 08, No: 01, Hal: 34-48: April. DOI: http://doi.org/10.51170/jii.v8i1.330