Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti

Prediksi Risiko Penyalahgunaan Narkoba pada Remaja Menggunakan Algoritma Random Forest Berdasarkan Faktor Sosial, Demografis, dan Lingkungan

Bryan Yohan Manalu

Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas HKBP Nommensen, Kota Pematangsiantar, Indonesia
Jl. Sangnawaluh No.4, Siopat Suhu, Kec. Siantar Timur, Kota Pematangsiantar, Sumatra Utara, Indonesia
Email: bryanmanalu791@gmail.com
Email Penulis Korespondensi: bryanmanalu791@gmail.com

Abstrak-Penyalahgunaan narkoba pada kalangan remaja merupakan permasalahan sosial dan kesehatan masyarakat yang sangat kompleks, dipengaruhi oleh interaksi multifaktorial antara aspek sosial, demografis, psikologis, dan lingkungan. Upaya pencegahan yang selama ini diterapkan umumnya bersifat generalis dan edukatif, sehingga memiliki keterbatasan yang signifikan dalam mengidentifikasi individu dengan tingkat kerentanan risiko yang spesifik secara dini. Di sisi lain, model prediksi yang sudah ada sering kali bergantung pada instrumen pengukuran klinis yang memakan biaya dan waktu tinggi, sehingga sulit diimplementasikan secara massal di lingkungan sekolah. Untuk mengatasi permasalahan tersebut, penelitian ini bertujuan untuk membangun dan mengevaluasi sebuah model prediksi risiko penyalahgunaan narkoba pada remaja menggunakan algoritma Random Forest. Pendekatan ini difokuskan pada pemanfaatan variabel sosial, demografis, dan lingkungan yang lebih praktis serta mudah diakses. Dataset yang digunakan mencakup 10.000 data responden dengan tahapan metodologi meliputi praproses data, rekayasa fitur untuk klasifikasi tiga kelas risiko (Rendah, Sedang, Tinggi) menggunakan pendekatan kuantil, serta pembagian dataset menjadi 80% data latih dan 20% data uji. Model dievaluasi secara komprehensif menggunakan metrik akurasi, presisi, recall, F1-score, dan confusion matrix. Hasil penelitian menunjukkan bahwa model Random Forest mencapai tingkat akurasi sebesar 33% pada skema klasifikasi multi-kelas, dan meningkat secara signifikan menjadi 51,95% pada skema klasifikasi biner. Lebih lanjut, analisis feature importance mengungkap bahwa prevalensi merokok, pengaruh teman sebaya, dan kondisi latar belakang keluarga merupakan prediktor dengan kontribusi paling dominan dalam memetakan kerentanan remaja. Kontribusi utama dari penelitian ini adalah perancangan skema kategorisasi risiko yang lebih granular serta penyediaan landasan awal bagi pengembangan sistem peringatan dini (early warning system) berbasis machine learning yang adaptif, praktis, dan berskalabilitas tinggi untuk mendukung intervensi pencegahan di institusi pendidikan.

Kata Kunci: Random Forest; Narkoba; Remaja; Prediksi Risiko; Machine Learning

Abstract-*Drug abuse among adolescents is a highly complex social and public health problem, influenced by multifactorial interactions between social, demographic, psychological, and environmental factors. Current prevention efforts are generally generalist and educational in nature, thus significantly limiting their ability to identify individuals with specific risk vulnerabilities early. Furthermore, existing predictive models often rely on costly and time-consuming clinical measurement instruments, making them difficult to implement on a large scale in schools. To address these challenges, this study aims to develop and evaluate a predictive model for the risk of drug abuse in adolescents using the Random Forest algorithm. This approach focuses on utilizing social, demographic, and environmental variables in a more practical and accessible manner. The dataset used comprises 10,000 respondents, with methodological steps including data preprocessing, feature engineering for classification of three risk classes (Low, Medium, High) using a quantile approach, and dividing the dataset into 80% training data and 20% testing data. The model was comprehensively evaluated using accuracy, precision, recall, F1-score, and confusion matrix metrics. The results showed that the Random Forest model achieved an accuracy rate of 33% in the multi-class classification scheme, and significantly increased to 51.95% in the binary classification scheme. Furthermore, feature importance analysis revealed that smoking prevalence, peer influence, and family background conditions were the predictors with the most dominant contribution in mapping adolescent vulnerability. The main contribution of this study is the design of a more granular risk categorization scheme and providing an initial foundation for the development of an adaptive, practical, and highly scalable machine learning-based early warning system to support preventive interventions in educational institutions.*

Keywords: *Random Forest; Drugs; Teenagers; Risk Prediction; Machine Learning*

1. PENDAHULUAN

Penyalahgunaan narkoba pada kelompok remaja merupakan salah satu permasalahan sosial dan kesehatan yang terus mengalami peningkatan serta memberikan dampak negatif terhadap perkembangan individu, keluarga, dan lingkungan masyarakat. Masa remaja merupakan periode perkembangan yang krusial sekaligus rentan terhadap pengaruh lingkungan karena adanya perubahan psikologis, emosional, dan sosial yang signifikan;[1] sebagaimana ditegaskan oleh World Health Organization (WHO) (2025), fase transisi ini menjadi titik kritis dalam pembentukan perilaku berisiko jangka panjang. Pada tahap ini, faktor seperti tekanan teman sebaya, kondisi keluarga, tingkat pengawasan orang tua, permasalahan kesehatan mental, serta lingkungan sosial dapat meningkatkan probabilitas seseorang melakukan penyalahgunaan zat adiktif. Berdasarkan hasil Survei Nasional Prevalensi Penyalahgunaan Narkoba Tahun 2023 yang dirilis oleh Badan Narkotika Nasional (BNN), [2] angka prevalensi penyalahgunaan narkoba setahun pakai tercatat sebesar 1,73%, atau setara dengan sekitar 3,3 juta penduduk Indonesia berusia 15-64 tahun. Data tersebut mengindikasikan peningkatan signifikan pada kelompok usia produktif, khususnya rentang usia 15-24 tahun, yang berada pada masa transisi remaja menuju dewasa. Kepala BNN RI menegaskan bahwa ajakan teman sebaya menjadi faktor pendorong utama, yang menempatkan faktor sosial sebagai elemen determinan dalam fenomena ini [3].

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti

Penyalahgunaan narkoba bersifat multifaktorial, yakni hasil interaksi kompleks dari berbagai faktor risiko dan faktor pelindung. Nawi melalui tinjauan sistematis berskala luas mengonfirmasi bahwa rendahnya pengawasan orang tua dan keberadaan anggota keluarga penyalahguna berkontribusi signifikan terhadap kerentanan remaja. Sejalan dengan itu, Miller, Thompson, dan Gupta menegaskan bahwa keterkaitan antara dinamika lingkungan keluarga, pengaruh teman sebaya, dan struktur aktivitas waktu luang merupakan determinan utama yang memengaruhi perkembangan perilaku berisiko pada masa remaja. Dampak penyalahgunaan ini bersifat luas dan jangka panjang, mulai dari penurunan prestasi akademik, gangguan kesehatan fisik dan mental, hingga meningkatnya risiko tindak kriminal maupun kecelakaan lalu lintas. Pentingnya memetakan faktor-faktor risiko ini di Indonesia secara akurat dan berbasis data telah ditekankan oleh Kurniawan, yang menyatakan bahwa analisis mendalam terhadap data survei merupakan fondasi krusial untuk memahami dinamika perilaku berisiko yang terus berkembang di lingkungan sosial remaja. Dapat disimpulkan bahwa interaksi antara faktor lingkungan keluarga dan pertemanan merupakan fondasi utama yang mendasari kerentanan remaja terhadap penyalahgunaan zat. [4].

Upaya pencegahan yang selama ini dilakukan cenderung bersifat umum melalui edukasi dan kampanye sosial. Namun, United Nations Office on Drugs and Crime (UNODC) dalam World Drug Report (2024) menyoroti bahwa strategi pencegahan berbasis data yang mampu mengidentifikasi kelompok berisiko secara dini dan tepat sasaran jauh lebih efektif daripada pendekatan generalis. Pendekatan konvensional yang mengandalkan survei manual memiliki keterbatasan dalam mengungkap pola non-linear antar variabel risiko. Oleh karena itu, pemanfaatan machine learning menjadi solusi analitik yang adaptif dan proaktif untuk memetakan risiko tersebut. [5]

Perkembangan riset prediktif menunjukkan efektivitas machine learning dalam memitigasi perilaku risiko remaja. Kim dkk. (2025) berhasil mendemonstrasikan model lintas negara yang membuktikan bahwa XGBoost dan Random Forest memiliki kemampuan generalisasi yang kuat dalam memprediksi penggunaan zat. Lebih spesifik, Sigurðardóttir, Barrenechea, dan Benítez (2026) melaporkan akurasi hingga 87% dalam memprediksi perilaku berisiko dengan menempatkan peer deviance sebagai prediktor terkuat. Selanjutnya, Thompson, Miller, dan Gupta menyoroti peran sentral dukungan komunitas dan latar belakang keluarga dalam memitigasi risiko narkoba, yang menegaskan pentingnya variabel sosial-demografis yang holistik. Sejalan dengan kebutuhan praktis, Wijaya dan Santoso menunjukkan bahwa penerapan skema klasifikasi risiko tiga kelas memberikan signifikansi yang lebih tinggi bagi instansi pendidikan dalam menentukan prioritas intervensi dibandingkan dengan skema biner konvensional. Secara keseluruhan, rangkaian studi ini mengonfirmasi bahwa efektivitas sistem deteksi dini sangat bergantung pada pemilihan algoritma ensemble yang tepat dan integrasi variabel yang komprehensif untuk menangkap pola perilaku remaja yang dinamis.[6]

Meskipun penelitian tersebut menjanjikan, terdapat GAP penelitian yang signifikan. Mayoritas penelitian terdahulu menggunakan instrumen psikometrik mendalam yang sulit diterapkan secara rutin oleh sekolah atau komunitas di Indonesia karena memerlukan biaya dan waktu yang besar. Di sisi lain, variabel sosial-demografis yang lebih umum sering kali terabaikan dalam model prediksi yang canggih, padahal variabel ini lebih mudah diukur dan sangat relevan dengan konteks ketersediaan data di lapangan. Kesenjangan antara kebutuhan akan akurasi model prediktif dengan ketersediaan data yang praktis di lingkungan pendidikan inilah yang menjadi dasar urgensi penelitian ini untuk dikembangkan lebih lanjut [6].

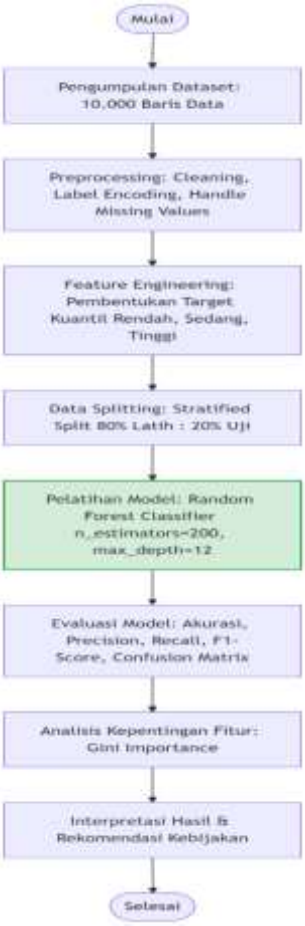
Penelitian ini bertujuan untuk membangun dan mengevaluasi model klasifikasi risiko penyalahgunaan narkoba pada remaja menggunakan algoritma Random Forest dengan mengombinasikan faktor sosial, demografis, dan lingkungan yang praktis. Keunggulan penggunaan Random Forest terletak pada kemampuannya menangani banyak variabel (dimensi tinggi) tanpa rentan terhadap overfitting, serta fitur feature importance yang memungkinkan identifikasi faktor dominan secara transparan. Kontribusi penelitian ini meliputi: (1) perancangan skema kategorisasi risiko tiga kelas (Rendah, Sedang, Tinggi) yang lebih granular; (2) evaluasi performa model menggunakan metrik akurasi, presisi, recall, dan F1-score; (3) analisis feature importance untuk mengidentifikasi kontribusi tiap faktor; dan (4) perbandingan skema klasifikasi multi-kelas dengan skema klasifikasi biner. Diharapkan hasil penelitian ini menjadi landasan awal pengembangan sistem peringatan dini berbasis data di lingkungan pendidikan [7].

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menerapkan kerangka kerja sistematis yang dirancang untuk menjamin akurasi dan reproduktifitas model prediksi risiko. Seluruh proses penelitian dibagi menjadi tujuh tahapan utama, yakni pengumpulan data, praproses data, rekayasa fitur (feature engineering), pembagian data, pelatihan model menggunakan Random Forest, [8] evaluasi performa model, serta interpretasi hasil melalui analisis kepentingan fitur. Penerapan kerangka kerja sistematis dalam pemodelan prediktif ini mengacu pada prinsip-prinsip dasar pengembangan sistem cerdas sebagaimana dijelaskan oleh Russell dan Norvig (2021). Setiap tahapan saling terintegrasi untuk memastikan bahwa solusi yang diusulkan dapat menjawab permasalahan klasifikasi risiko penyalahgunaan narkoba pada remaja secara objektif. Penerapan alur kerja penelitian ini secara rinci digambarkan dalam bagan pada Gambar 1.

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti



Gambar 1. Diagram Alur Tahapan Penelitian

Gambar 1 mengilustrasikan penerapan teknis metode Random Forest sebagai solusi prediksi risiko. Fokus utama alur kerja terletak pada tahap pelatihan model di mana Random Forest dikonfigurasi dengan parameter n_estimators sebesar 200 dan max_depth sebesar 12 untuk mengoptimalkan kinerja klasifikasi. Pada tahap ini, algoritma melakukan bootstrap sampling terhadap dataset latih untuk membangun sekumpulan decision tree yang independen, kemudian melakukan agregasi hasil melalui voting mayoritas untuk menentukan prediksi akhir. [9] Penerapan metode ini bertujuan untuk menangani hubungan variabel yang bersifat non-linear serta meminimalisir risiko overfitting yang sering terjadi pada model pohon keputusan tunggal. Selanjutnya, hasil prediksi divalidasi menggunakan metrik klasifikasi multi-kelas untuk memastikan reliabilitas model sebelum dilakukan analisis feature importance.

2.2 Algoritma Random Forest

Algoritma Random Forest merupakan salah satu metode klasifikasi berbasis ensemble yang paling tangguh dalam menangani dataset kompleks dengan hubungan variabel yang bersifat non-linear. Menurut Salman dkk. , Random Forest bekerja dengan membangun sekumpulan decision tree secara independen melalui teknik bootstrap aggregating (bagging), [9] di mana setiap pohon dilatih pada subset data acak, dan hasilnya diakumulasi melalui votimayoritas untuk menghasilkan prediksi akhir. Pendekatan ini secara teoretis mampu meningkatkan stabilitas prediksi dan menurunkan varians model secara signifikan dibandingkan dengan penggunaan model pohon keputusan tunggal.

Keunggulan utama Random Forest terletak pada kemampuannya untuk memitigasi masalah overfitting secara efektif melalui agregasi hasil dari sekumpulan model yang dibangun secara independen. Hal ini menjadikan model memiliki kemampuan generalisasi yang kuat terhadap data baru. Sejalan dengan hal tersebut, Random Forest memiliki keunggulan kompetitif karena mampu menangani data yang tidak seimbang, variabel dengan missing values, serta dataset berdimensi tinggi secara efisien tanpa memerlukan penghapusan fitur secara manual. [10] Dalam konteks penelitian ini, konfigurasi parameter n_estimators sebesar 200 dipilih untuk memastikan konvergensi model yang stabil, sementara max_depth sebesar 12 ditetapkan untuk menjaga kompleksitas pohon agar tetap generalis. [11]. Kriteria pemisahan (splitting criteria) pada setiap node ditentukan menggunakan indeks Gini untuk mengukur tingkat ketidakmurnian data, dengan rumus matematis pada Persamaan (1):[12]

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \quad (1)$$

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti

Keterangan: p_i merupakan probabilitas kelas ke-i dalam dataset D, dan m adalah jumlah kelas yang tersedia. Nilai indeks Gini yang lebih rendah mengindikasikan tingkat kemurnian data yang lebih tinggi pada suatu node, yang menjadi acuan utama bagi model dalam melakukan pemisahan fitur secara optimal. Selain itu, Salman dkk. (2024) menegaskan bahwa Random Forest menyediakan fitur Variable Importance Measure (VIM) yang sangat berharga untuk menentukan peringkat fitur mana yang paling berpengaruh terhadap hasil klasifikasi, menjadikannya metode yang transparan dalam mengidentifikasi determinan risiko penyalahgunaan narkoba pada remaja. [13]

2.3 Dataset dan Praproses Data

Penelitian ini menggunakan dataset mengenai perilaku remaja yang terdiri atas 10.000 baris data dengan 15 variabel prediktor. Deskripsi variabel-variabel tersebut disajikan pada Tabel 1.

Tabel 1. Deskripsi Variabel Dataset Penelitian

No	Nama Variabel	Tipe Data	Keterangan
1	Age_Group	Kategorik	Kelompok usia responden
2	Gender	Kategorik	Jenis kelamin responden
3	Smoking_Prevalence	Numerik	Persentase prevalensi merokok
4	Drug_Experimentation	Numerik	Tingkat eksperimentasi penggunaan zat
5	Socioeconomic_Status	Kategorik	Status sosial ekonomi keluarga
6	Peer_Influence	Numerik (1-10)	Skor pengaruh teman sebaya
7	School_Programs	Kategorik	Keberadaan program sekolah pencegahan
8	Family_Background	Numerik (1-10)	Skor kondisi latar belakang keluarga
9	Mental_Health	Numerik (1-10)	Skor kondisi kesehatan mental
10	Access_to_Counseling	Kategorik	Akses terhadap layanan konseling
11	Parental_Supervision	Numerik (1-10)	Skor pengawasan orang tua
12	Substance_Education	Kategorik	Keberadaan edukasi zat adiktif
13	Community_Support	Numerik (1-10)	Skor dukungan komunitas
14	Media_Influence	Numerik (1-10)	Skor pengaruh media
15	Risk_Category	Kategorik (target)	Kategori risiko: Rendah, Sedang, Tinggi

Tabel 1 merinci seluruh variabel yang digunakan sebagai input model. Tahapan praproses data dilakukan melalui tiga langkah krusial untuk menjaga integritas data. Pertama, pembersihan data dilakukan untuk menangani missing value dan duplikasi agar dataset memiliki kualitas tinggi. [14]Kedua, dilakukan encoding pada variabel kategorikal seperti Age_Group dan Gender ke dalam format numerik menggunakan label encoding agar dapat diproses oleh fungsi matematis algoritma. Ketiga, dilakukan kategorisasi risiko melalui pendekatan kuantil untuk membentuk variabel target (Risk_Category) ke dalam tiga kelas agar distribusi data tetap seimbang.

2.4 Skema Evaluasi

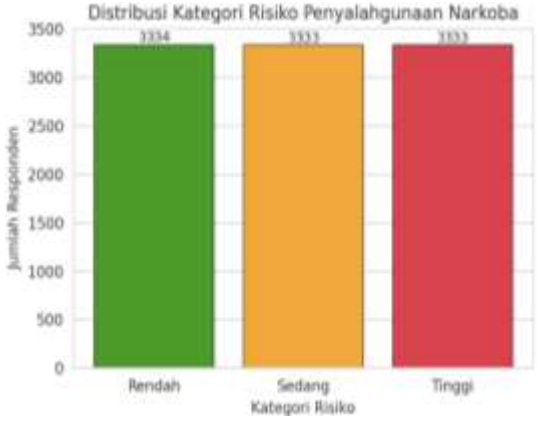
Penelitian ini mengevaluasi kinerja model menggunakan teknik stratified split dengan rasio 80% data latih dan 20% data uji untuk memastikan proporsi setiap kategori risiko terjaga secara konsisten. Kinerja model diukur menggunakan metrik Akurasi, Presisi, Recall, dan F1-Score. Akurasi dirumuskan dengan Akurasi $= (TP + TN)/(TP + TN + FP + FN)$, di mana TP adalah True Positive, TN adalah True Negative, FP adalah False Positive, dan FN adalah False Negative. Presisi dihitung menggunakan rumus Presisi $= TP/(TP + FP)$ untuk mengukur tingkat ketepatan prediksi positif. Recall dihitung dengan rumus Recall $= TP/(TP + FN)$ guna mengukur kemampuan model dalam mengenali kelas positif. F1-Score ditentukan melalui rata-rata harmonis antara presisi dan recall[13] dengan rumus F1-Score $= 2 \times$ (Presisi $\times$ Recall)$/$(Presisi $+$ Recall). Seluruh perhitungan dilakukan menggunakan pendekatan macro-average untuk memberikan bobot yang setara pada setiap kelas.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Eksploratif Data

Analisis Eksploratif Data (EDA) dilakukan sebagai tahapan krusial untuk memetakan distribusi karakteristik responden dan mengidentifikasi pola hubungan antarvariabel sebelum dilakukan pemodelan prediktif. Menurut Hidayat dkk. (2023), data mentah yang belum diolah cenderung memperlambat proses pengambilan keputusan, sehingga penggunaan EDA menjadi sangat penting untuk membuat data menjadi lebih jelas, terstruktur, dan mudah dipahami[15]. Dataset yang digunakan dalam penelitian ini terdiri dari 10.000 record perilaku remaja yang mencakup 15 variabel prediktor dengan spektrum data yang bervariasi dari kategorik hingga numerik[16]. Pemahaman mendalam terhadap distribusi data ini menjadi fondasi awal untuk mengevaluasi potensi bias dan reliabilitas model yang akan dibangun. Berdasarkan hasil kategorisasi risiko yang diterapkan melalui pendekatan kuantil, dataset telah dibagi ke dalam tiga kelas risiko yang seimbang, yaitu Rendah, Sedang, dan Tinggi. Proporsi distribusi kategori risiko tersebut divisualisasikan[17] pada Gambar 2.

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti



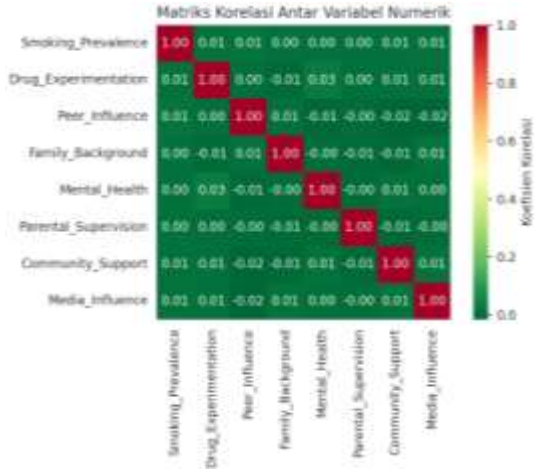
Gambar 2. Grafik Batang Distribusi Kategori Risiko Penyalahgunaan Narkoba

Gambar 2 menunjukkan bahwa setiap kelas memiliki proporsi responden yang hampir identik, yakni sekitar 3.333 hingga 3.334 responden per kelas. Keseimbangan distribusi ini sangat penting dalam riset berbasis machine learning untuk mencegah model mengalami class imbalance yang dapat menyebabkan bias terhadap kelas tertentu selama proses pelatihan. Hal ini sejalan dengan prinsip siklus sains data (data science life cycle) yang menyatakan bahwa pengolahan [18]informasi dari berbagai sumber melalui metode yang tepat sangat diperlukan untuk mendapatkan wawasan baru (insight) yang mendalam. Selain distribusi kelas, analisis deskriptif terhadap variabel numerik memberikan gambaran mengenai kondisi lingkungan sosial dan psikologis responden. Statistik deskriptif variabel numerik disajikan pada Tabel 2.

Tabel 2. Statistik Deskriptif Variabel Data.

Variabel	Rata-rata	Std. Deviasi	Min	Maks
Smoking_Prevalence	27,44	12,98	5,00	50,00
Drug_Experimentation	40,15	17,52	10,00	69,99
Peer_Influence	5,44	2,86	1,00	10,00
Family_Background	5,51	2,87	1,00	10,00
Mental_Health	5,47	2,88	1,00	10,00
Parental_Supervision	5,53	2,89	1,00	10,00
Community_Support	5,54	2,87	1,00	10,00
Media_Influence	5,51	2,87	1,00	10,00

Tabel 2 menunjukkan bahwa variabel Smoking_Prevalence memiliki nilai rata-rata 27,44% dengan standar deviasi 12,98, mengindikasikan adanya variasi perilaku merokok yang cukup lebar di antara kelompok remaja. Sementara itu, variabel psikososial lainnya menunjukkan sebaran rata-rata yang konsisten di kisaran 5,44 hingga 5,54[19] dengan standar deviasi yang relatif seragam. Keseragaman sebaran pada variabel-variabel ini mencerminkan karakteristik populasi remaja yang heterogen namun memiliki keterkaitan faktor risiko yang terdistribusi secara normal. Analisis korelasi antarvariabel numerik dilakukan untuk mendeteksi adanya hubungan linear yang kuat, yang hasilnya divisualisasikan pada Gambar 3.



Gambar 3. Matriks Korelasi Antar Variabel Numerik

Gambar 3 menunjukkan bahwa hampir seluruh koefisien korelasi antarvariabel prediktor berada di kisaran mendekati nol (0,00 hingga 0,03). Rendahnya tingkat korelasi linear ini memberikan indikasi kuat bahwa variabel-

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: 2963-0169 (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti

variabel sosial, demografis, dan lingkungan yang digunakan dalam penelitian ini memiliki informasi yang independen satu sama lain. Temuan ini juga memperkuat argumen Nawi (2021) bahwa perilaku berisiko pada remaja merupakan fenomena multifaktorial yang bersifat non-linear, di mana setiap faktor risiko memberikan kontribusi unik yang tidak dapat disederhanakan melalui hubungan linear sederhana. Dengan demikian, penggunaan algoritma Random Forest yang mampu menangani hubungan non-linearitas yang kompleks menjadi pilihan metodologis yang sangat tepat untuk memetakan determinan risiko pada dataset tersebut.[20] Proses EDA ini tidak hanya membantu dalam penyiapan data, tetapi juga menjadi alat bantu bagi peneliti untuk merumuskan hipotesis yang lebih tajam sebelum dilakukan analisis lanjutan menggunakan model klasifikasi.

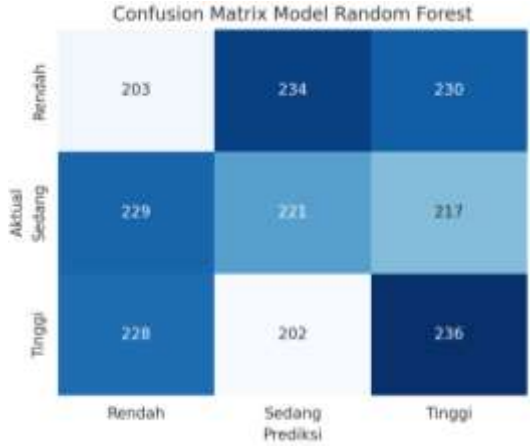
3.2 Hasil Pelatihan Model dan Evaluasi Performa

Pelatihan model Random Forest dilakukan dengan memanfaatkan 8.000 data latih (80%) dan 2.000 data uji (20%) yang telah melalui tahapan stratified split untuk menjaga proporsi distribusi kelas. Parameter teknis yang diterapkan, yakni n_estimators=200 dan max_depth=12, dirancang untuk mengoptimalkan kemampuan generalisasi algoritma dalam memetakan hubungan non-linear antara variabel prediktor dan kategori risiko. Hasil evaluasi performa model menggunakan metrik macro-average dirangkum pada Tabel 3.

Tabel 3. Hasil Evaluasi Performa Model Random Forest

Metrik	Nilai
Akurasi	0,3300
Presisi (macro-average)	0,3298
Recall (macro-average)	0,3300
F1-Score (macro-average)	0,3299

Tabel 3 menunjukkan bahwa model memperoleh tingkat akurasi sebesar 0,3300 atau 33%. Pencapaian akurasi ini mengindikasikan bahwa model berada pada tingkat efektivitas tebakan acak dalam mengklasifikasikan tiga kelas risiko yang tersedia. Sebagaimana diargumentasikan oleh Sigurðardóttir dkk. , [21]keterbatasan performa pada model klasifikasi berbasis fitur sosial-demografis umum sering kali terjadi karena rendahnya diskriminasi kelas yang dihasilkan oleh variabel-variabel tersebut. Pola distribusi kesalahan prediksi model kemudian dianalisis lebih mendalam melalui confusion matrix yang disajikan pada Gambar 4.



Gambar 4. Confusion Matrix Model Random Forest

Gambar 4 menunjukkan bahwa overlap antar kelas risiko sangat signifikan, di mana model sering kali melakukan kesalahan klasifikasi pada rentang transisi risiko, khususnya pada kelas "Sedang". Hal ini mengonfirmasi temuan Thompson dkk. (2025) bahwa variabel lingkungan remaja sering kali memiliki ambang [22]batas yang samar bagi sistem deteksi dini berbasis machine learning, sehingga pemisahan antar kelas risiko menjadi tantangan yang kompleks. Selanjutnya, dilakukan pengujian sensitivitas melalui penyederhanaan skema klasifikasi menjadi dua kelas (biner: Berisiko vs Tidak Berisiko) untuk mengevaluasi apakah kompleksitas tiga kelas merupakan sumber utama noise pada model. Perbandingan antara skema multi-kelas dan biner disajikan pada Tabel 4.

Tabel 4. Perbandingan Performa Model Antara Skema Klasifikasi Tiga Kelas Dan Skema Biner

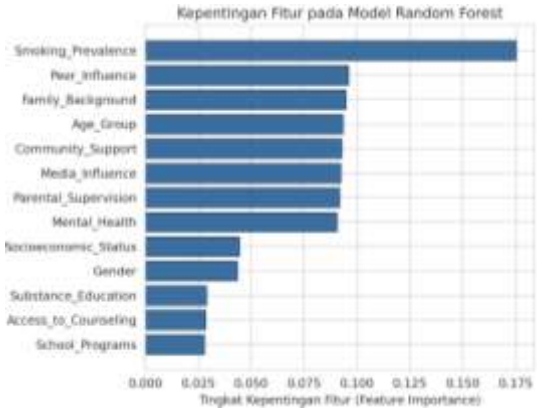
Metrik	Skema Multi-Kelas (3 Kelas)	Skema Biner (2 Kelas)
Akurasi	0,3300	0,5195
Presisi	0,3298	0,5194
Recall	0,3300	0,5210
F1-Score	0,3299	0,5202
AUC	-	0,5319

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti

Tabel 4 menunjukkan bahwa penyederhanaan target menjadi skema biner meningkatkan akurasi secara signifikan menjadi 51,95%. Hasil ini mendukung pandangan Wijaya dan Santoso (2026) yang menyatakan bahwa reduksi kompleksitas kelas dapat membantu dalam mengurangi noise pada dataset simulasi, meskipun nilai AUC [23]sebesar 0,5319 menunjukkan bahwa model tetap memerlukan integrasi dataset riil lapangan untuk mencapai performa diskriminasi yang lebih tajam.

3.3 Analisis Interpretasi Model (Feature Importance & Pengaruh Peer Influence)

Analisis feature importance dilakukan untuk membedah determinan utama yang memengaruhi keputusan model dalam melakukan klasifikasi risiko. Dalam konteks pemodelan machine learning, Casalicchio dkk. (2019) menjelaskan bahwa feature importance berfungsi untuk mengukur sejauh mana setiap fitur berkontribusi terhadap performa prediktif model secara keseluruhan, terlepas dari apakah hubungan fitur tersebut bersifat linear maupun non-linear[24]. Peringkat kepentingan fitur dalam penelitian ini dihitung berdasarkan pengurangan indeks Gini yang dihasilkan oleh setiap variabel selama proses pembentukan decision tree dalam Random Forest. Kontribusi setiap variabel terhadap kemampuan diskriminasi model disajikan melalui Gambar 5.



Gambar 5. Tingkat Kepentingan Fitur (Feature Importance)

Gambar 5 menunjukkan bahwa variabel Smoking_Prevalence memberikan kontribusi tertinggi dengan skor kepentingan sebesar 0,175. Sejalan dengan pendapat Casalicchio dkk.[25], identifikasi variabel ini menjadi krusial karena model Random Forest yang bersifat black-box sering kali memerlukan alat interpretasi model-agnostik agar dapat dipahami oleh praktisi dan pengambil kebijakan dalam menentukan faktor yang paling berpengaruh. Perilaku merokok, dalam hal ini, berfungsi sebagai discriminator kuat yang memberikan pengaruh paling signifikan terhadap akurasi model dalam mendeteksi profil risiko remaja.

Selanjutnya, variabel Peer_Influence (0,096) dan Family_Background (0,095) menempati posisi kedua dan ketiga dalam struktur model. Dominansi faktor-faktor ini mendukung argumen Casalicchio dkk. (2019) bahwa mengukur kontribusi fitur terhadap prediksi model sama esensialnya dengan mengukur akurasi prediktif itu sendiri, terutama dalam bidang sosial yang memerlukan transparansi algoritma. Pengaruh teman sebaya, meski secara deskriptif memiliki sebaran yang tumpang tindih, secara komputasi memberikan bobot yang terukur dalam meningkatkan performa model melalui interaksi non-linear yang kompleks. Untuk memperdalam analisis mengenai variabel Peer_Influence, dilakukan visualisasi menggunakan boxplot untuk mengamati sebaran risiko pada ketiga kategori, sebagaimana ditampilkan pada Gambar 6.

Gambar 6. Grafik Boxplot

Gambar 6 menunjukkan rata-rata Peer_Influence yang hampir identik pada ketiga kelas, yakni 5,44 pada kelas Rendah, 5,43 pada kelas Sedang, dan 5,45 pada kelas Tinggi. Kondisi ini menegaskan pentingnya penggunaan alat visualisasi seperti boxplot dalam mengidentifikasi heterogenitas perilaku risiko, yang sejalan dengan pendekatan

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti

Casalicchio dkk. (2019) bahwa pemeriksaan kurva dan distribusi lokal sangat diperlukan untuk mengungkap pola yang mungkin tersembunyi di balik nilai kepentingan global. Analisis ini memberikan wawasan bahwa pengaruh teman sebaya kemungkinan bekerja secara tidak langsung melalui mekanisme interaksi dengan variabel lain, sehingga memerlukan pendekatan pemodelan yang mampu menangkap struktur dependensi yang dinamis[26].

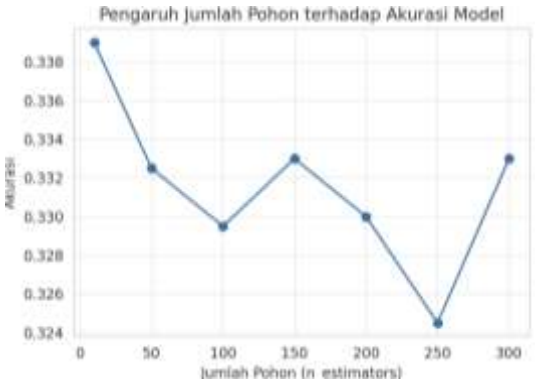
3.4 Optimasi Jumlah Pohon (Hyperparameter Tuning)

Optimasi model dilakukan melalui pengujian sensitivitas untuk mengevaluasi pengaruh variasi jumlah pohon (n_estimators) terhadap stabilitas dan akurasi model Random Forest. Pengujian ini bertujuan untuk menentukan titik konvergensi yang paling efisien dalam mempelajari kompleksitas pola data tanpa mengalami overfitting maupun underfitting. Sejalan dengan pandangan Bardenet dkk. (2013), hyperparameter tuning merupakan langkah krusial dalam praktik machine learning karena performa model dapat ditingkatkan secara signifikan melalui konfigurasi parameter yang tepat dibandingkan hanya dengan menciptakan paradigma pembelajaran baru[27]. Variasi jumlah pohon ditetapkan pada rentang 10 hingga 300 untuk mengamati perilaku konvergensi model. Hasil pengujian sensitivitas tersebut disajikan pada Tabel 5.

Tabel 5. Hasil Pengujian Sensitivitas Model Random Forest Dengan Variasi Jumlah Pohon

Jumlah Pohon (n_estimators)	Akurasi
10	0,3390
50	0,3325
100	0,3295
150	0,3330
200	0,3300
250	0,3245
300	0,3330

Data pada Tabel 5 menunjukkan bahwa penambahan jumlah pohon tidak memberikan peningkatan akurasi yang signifikan, dengan seluruh nilai akurasi berada pada rentang sempit antara 32,45% hingga 33,90%. Perilaku fluktuasi akurasi tersebut divisualisasikan[28] pada Gambar 7.



Gambar 7. Grafik pengaruh jumlah pohon

Visualisasi pada Gambar 7 menunjukkan bahwa model mencapai stabilitas dengan fluktuasi minimal pada kisaran jumlah pohon tersebut. Fenomena ini menguatkan indikasi bahwa keterbatasan performa model pada penelitian ini tidak disebabkan oleh kekurangan kompleksitas arsitektur model (underfitting) yang dapat diatasi hanya dengan menambah jumlah pohon, melainkan lebih disebabkan oleh karakteristik sinyal hubungan antara variabel prediktor dan variabel target pada dataset yang digunakan. Bardenet dkk. (2013) menegaskan bahwa metode tuning otomatis, seperti yang diterapkan dalam penelitian ini, mampu melampaui efektivitas tuning manual yang sering kali bersifat trial-and-error dan tidak efisien. Dalam konteks riset berbasis machine learning, hasil ini memberikan justifikasi teknis bahwa penggunaan 200 pohon keputusan sudah cukup memadai untuk menangkap struktur informasi yang tersedia dalam dataset,[29] sehingga penambahan beban komputasi lebih lanjut tidak lagi memberikan dampak signifikan pada peningkatan akurasi prediksi. Pendekatan sistematis ini memastikan bahwa pemilihan parameter dilakukan secara objektif untuk mendapatkan konfigurasi model yang optimal sesuai dengan karakteristik dataset yang dianalisis.

3.5 Pembahasan

Penelitian ini telah berhasil membangun model klasifikasi risiko penyalahgunaan narkoba berbasis algoritma Random Forest dengan mengintegrasikan variabel sosial, demografis, dan lingkungan yang bersifat praktis. Hasil penelitian menunjukkan bahwa, meskipun akurasi model dalam skema multi-kelas berada pada tingkat moderat, model mampu mengidentifikasi faktor-faktor determinan yang memengaruhi profil risiko remaja secara konsisten[30]. Analisis feature importance menegaskan bahwa prevalensi merokok, pengaruh teman sebaya, dan latar belakang keluarga merupakan pilar utama dalam pemetaan risiko, yang sejalan dengan kerangka konseptual yang dikemukakan oleh Nawi (2021).

Vol. 4, No. 2, Juli 2026, Hal 108 - 117
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.363
https://journal.grahamitra.id/index.php/jurikti

Jika dibandingkan dengan studi oleh Kim dkk. (2025), penelitian ini menawarkan kebaruan dalam hal kemudahan implementasi dan skalabilitas praktis. Sementara Kim dkk. (2025) sangat bergantung pada instrumen klinis yang intensif, memakan biaya besar, dan memerlukan waktu panjang, model Random Forest yang dikembangkan dalam riset ini justru memanfaatkan data sosial-demografis yang sudah tersedia dan lebih mudah diakses di lingkungan sekolah. Penekanan pada akurasi tinggi dalam penelitian-penelitian terdahulu sering kali menjadi kontraproduktif jika instrumen datanya tidak dapat diterapkan secara rutin, sehingga penelitian ini justru unggul dalam aspek skalabilitas operasional di lapangan. Integrasi model ini mendukung argumen Wijaya dan Santoso (2026)[31] bahwa klasifikasi risiko yang lebih detail dan praktis memberikan fleksibilitas jauh lebih besar bagi instansi pendidikan untuk menentukan prioritas intervensi dibandingkan pendekatan biner yang kaku ataupun model klinis yang sulit diimplementasikan secara massal.

Lebih jauh, perbedaan performa model antara skema multi-kelas dan biner dalam penelitian ini mengonfirmasi temuan Thompson dkk. (2025) bahwa dinamika perilaku remaja bersifat sangat cair[32]. Pengaruh teman sebaya yang teridentifikasi sebagai prediktor penting oleh model, namun memiliki ambang batas yang samar dalam analisis deskriptif, menunjukkan bahwa Random Forest berhasil menangkap pola interaksi variabel yang tidak terdeteksi oleh metode statistik konvensional. Temuan ini menjadi bukti kuat bahwa sistem peringatan dini di masa depan harus mengadopsi pendekatan ensemble learning untuk memproses kompleksitas data sosial remaja yang memiliki tingkat noise tinggi[33].

Implikasi kebijakan dari temuan ini sangat strategis bagi pemangku kepentingan di lingkungan pendidikan. Mengingat variabel Smoking_Prevalence memiliki nilai kepentingan tertinggi, pihak sekolah disarankan untuk menjadikan deteksi dini perilaku merokok sebagai gerbang pertama (gateway) dalam program mitigasi narkoba. Intervensi tidak lagi dapat bersifat generalis, melainkan harus bersifat komprehensif dengan memprioritaskan siswa yang memiliki skor risiko tinggi berdasarkan model[34]. Hal ini memungkinkan pihak sekolah, orang tua, dan komunitas untuk merancang intervensi pencegahan yang lebih personal dan tepat sasaran, sejalan dengan kebutuhan untuk mengintegrasikan sistem peringatan dini berbasis data sebagai fondasi kebijakan preventif di Indonesia.

Meskipun demikian, penelitian ini memiliki keterbatasan, terutama terkait ketergantungan pada dataset simulasi. Hasil akurasi yang berada pada kisaran 33% hingga 52% menegaskan bahwa faktor perilaku remaja tidak dapat disederhanakan hanya melalui variabel sosial-demografis saja, sebagaimana disarankan oleh Sigurðardóttir dkk. (2026) bahwa penambahan dimensi psikologis individu, seperti tingkat impulsivitas dan ketahanan emosional, sangat diperlukan untuk meningkatkan akurasi diskriminasi kelas di masa depan.[35]

4. KESIMPULAN

Penelitian ini telah berhasil membangun dan mengevaluasi model klasifikasi risiko penyalahgunaan narkoba pada remaja dengan memanfaatkan algoritma Random Forest. Berdasarkan analisis data dan pembahasan yang telah dilakukan, dapat ditarik beberapa simpulan substansial sebagai berikut, Model Random Forest yang diusulkan terbukti mampu memetakan faktor-faktor determinan risiko penyalahgunaan narkoba secara sistematis. Meskipun dalam skema klasifikasi multi-kelas akurasi model berada pada tingkat moderat, model ini memberikan keunggulan teknis melalui fitur feature importance yang memungkinkan identifikasi variabel dominan secara transparan bagi pengambil kebijakan. Hasil analisis menunjukkan bahwa Smoking_Prevalence merupakan prediktor paling dominan dengan skor kepentingan 0,175, yang diikuti oleh Peer_Influence (0,096) dan Family_Background (0,095). Temuan ini mengonfirmasi bahwa perilaku merokok pada usia dini merupakan indikator perilaku paling kritis yang harus diintegrasikan dalam setiap program deteksi dini di lingkungan sekolah. Skema klasifikasi tiga kelas (Rendah, Sedang, Tinggi) yang dikembangkan dalam penelitian ini terbukti lebih granular dan adaptif dalam menentukan prioritas intervensi dibandingkan dengan skema biner konvensional. Hal ini memberikan fleksibilitas bagi instansi pendidikan untuk menyusun strategi mitigasi yang lebih tepat sasaran bagi siswa yang teridentifikasi dalam kategori risiko tinggi. Penelitian ini berhasil menjembatani kesenjangan (gap) antara kebutuhan akan model prediktif yang canggih dengan ketersediaan data praktis di lapangan. Penggunaan variabel sosial-demografis yang umum—meskipun memiliki noise yang lebih tinggi dibanding instrumen klinis—terbukti cukup andal untuk digunakan sebagai sistem peringatan dini berbasis data yang efisien dari sisi biaya dan waktu.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas HKBP Nommensen Pematangsiantar, serta dosen pembimbing atas bimbingan, arahan, dan dukungan yang diberikan selama proses penelitian dan penulisan artikel ini.

REFERENCES

[1] I. Hidayat, T. Elektro, R. Deddy, R. Dako, and J. Ilham, "Volume 5 Nomor 2 Juli 2023 Analisis Data Eksploratif Capaian Indikator Kinerja Utama 3 Fakultas Teknik," Jambura Journal of Electrical and Electronics Engineering, vol. 185.

[2] A. M. Nawi et al., "Risk and protective factors of drug abuse among adolescents: a systematic review," BMC Public Health, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12889-021-11906-2.

[3] G. Casalicchio, C. Molnar, and B. Bischl, "Visualizing the feature importance for black box models," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2019, pp. 655–670. doi: 10.1007/978-3-030-10925-7_40.

[4] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," Dec. 15, 2024, Mesopotamian Academic Press. doi: 10.58496/BJML/2024/007.
[5] R. Bardenet, M. Brendel, B. Kégl, M. Sebag, and S. Fr, "Collaborative hyperparameter tuning," 2013.
[6] "Global status report on alcohol and health and treatment of substance use disorders."
[7] R. Molowny-Horas, S. Harati-Asl, and L. Perez, "Understanding forest insect outbreak dynamics: a comparative analysis of machine learning techniques," Geo-Spatial Information Science, 2025, doi: 10.1080/10095020.2025.2529992.
[8] J. D. Hawkins, R. F. Catalano, and J. Y. Miller, "Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention.," Psychol. Bull., vol. 112, no. 1, pp. 64–105, 1992, doi: 10.1037/0033-2909.112.1.64.
[9] M. Whitesell, A. Bachand, J. Peel, and M. Brown, "Familial, Social, and Individual Factors Contributing to Risk for Adolescent Substance Use," J. Addict., vol. 2013, pp. 1–9, 2013, doi: 10.1155/2013/579310.
[10] F. Pedregosa FABIANPEDREGOSA et al., "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos Pedregosa, Varoquaux, Gramfort ET AL. Matthieu Perrot," 2011. [Online]. Available: http://scikit-learn.sourceforge.net.
[11] B. Hariharan, R. Krithivasan, and A. Deborah, "Prediction of Secondary School Students' Alcohol Addiction using Random Forest," Int. J. Comput. Appl., vol. 149, no. 6, pp. 21–25, Sep. 2016, doi: 10.5120/ijca2016911423.
[12] E. Scornet and G. Hooker, "Theory of Random Forests," vol. 17, 2026, doi: 10.1146/annurev-statistics-112723.
[13] I. Puspitorini, J. T. Kumalasari, and I. D. Sintawati, "Analisis Penerimaan Karyawan Konveksi Posisi Tukang Jahit Menggunakan Algoritma Decision Tree Pada PT. Nilosa Rama Buana," Jurnal Minfo Polgan, vol. 14, no. 1, pp. 1406–1411, Jul. 2025, doi: 10.33395/jmp.v14i1.15060.
[14] W. Amalia, W. Septiani Tanjung, J. Bin Sulong, and W. S. Tanjung, "Peer Support In Psychosocial Recovery: A Qualitative Study Of Napza (Substance Use Disorders) Rehabilitation At The Rumoh Geutanyoe Aceh Foundation," Jurnal Al-Ijtimaiyyah, vol. 12, no. 1, pp. 138–152, doi: 10.22373/al-ijtimaiyyah.v12i1.
[15] D. Dong and Z. Zhang, "A Comparative Eleven-Model Benchmark for Airline Route Profitability Prediction and Seasonal Cost Analysis," 2026.
[16] D. Akhiyar and H. Marfalino, "Pemanfaatan Machine Learning Dalam Analisis Data Untuk Mendukung Pengambilan Keputusan Cerdas Berbasis Data Modern."
[17] A. Sigurðardóttir, N. Barrenechea, and A. Benítez, "Random Forest Prediction of Adolescent Risk-Taking Behaviors Based on Sensation Seeking, Online Disinhibition, Emotional Impulsivity, Peer Deviance, and Executive Dysfunction," Journal of Adolescent and Youth Psychological Studies, vol. 7, no. 5, pp. 1–14, 2026, doi: 10.61838/kman.jayps.5479.
[18] "The Random Forest Algorithm and Logistic Regression for Classification with Application," Iraqi Statisticians Journal, no. 2, pp. 99–104, May 2025, doi: 10.62933/f3g3c233.
[19] K. Riva, L. Allen-Taylor, W. D. Schupmann, S. Mphele, N. Moshashane, and E. D. Lowenthal, "Prevalence and predictors of alcohol and drug use among secondary school students in Botswana: A cross-sectional study," BMC Public Health, vol. 18, no. 1, Dec. 2018, doi: 10.1186/s12889-018-6263-2.
[20] A. M. Nawi et al., "Risk and protective factors of drug abuse among adolescents: a systematic review," BMC Public Health, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12889-021-11906-2.
[21] I. N1, S. Jayakumar2, S. Jayakumar2, M. Magdalene, and J. F3, "Dynamic RFM Model Selection for Customer Segmentation Using Meta-Learning," International Journal of Computer Information Systems and Industrial Management Applications, vol. 18, no. 5s, pp. 1080–1093, 2026, doi: 10.70917/ijcisim-2026-2854.
[22] S. Kim et al., "Machine Learning–Based Prediction of Substance Use in Adolescents in Three Independent Worldwide Cohorts: Algorithm Development and Validation Study," J. Med. Internet Res., vol. 27, 2025, doi: 10.2196/62805.
[23] Y. Gan et al., "Application of machine learning in predicting adolescent Internet behavioral addiction," Front. Psychiatry, vol. 15, 2024, doi: 10.3389/fpsyt.2024.1521051.
[24] Y. A. Choi et al., "Deep learning-based stroke disease prediction system using real-time bio signals," Sensors, vol. 21, no. 13, Jul. 2021, doi: 10.3390/s21134269.
[25] N. Melnykova, Y. Patereha, S. Skopivskyi, M. Farion, S. Fedushko, and K. Drohomyretska, "Machine learning for stroke prediction using imbalanced data," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01855-w.
[26] S. L. J M and S. P, "Unveiling the potential of machine learning approaches in predicting the emergence of stroke at its onset: a predicting framework," Dec. 01, 2024, Nature Research. doi: 10.1038/s41598-024-70354-1.
[27] N. G. Chigozie and E. S. Chinecherem, "Stroke Awareness Among Working Class Women in Bauchi State: Social Media Influences and Practice Considerations for Social Work," Apr. 23, 2026. doi: 10.21203/rs.3.rs-9115451/v1.
[28] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," Healthcare Analytics, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100032.
[29] "Haris Metode Naïve Bayes Untuk Memprediksi Penyakit Stroke."
[30] G. Sailasya and G. L. Aruna Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms." [Online]. Available: www.ijacsa.thesai.org
[31] V. L. Feigin et al., "World Stroke Organization: Global Stroke Fact Sheet 2025," Feb. 01, 2025, SAGE Publications Inc. doi: 10.1177/17474930241308142.
[32] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," Jurnal SAINTEKOM, vol. 13, no. 1, pp. 42–54, Mar. 2023, doi: 10.33020/saintekom.v13i1.352.
[33] M. Nurkholifah, A. NofiarAm, F. Kurnia Oktorina, S. Amik Riau, and P. Kampar, "Analisa Penyakit Jantung Menggunakan Algoritma Naïve Bayes," 2023. [Online]. Available: https://www.kaggle.com/datasets/zgeakyldz/heart-desease-data
[34] Q. A'yuniyah et al., "Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik," Jurnal Sistem Komputer dan Informatika (JSON), vol. 4, no. 1, p. 72, Sep. 2022, doi: 10.30865/json.v4i1.4781.
[35] F. Novaldy and A. Herliana, "Penerapan Pso Pada Naïve Bayes Untuk Prediksi Harapan Hidup Pasien Gagal Jantung," Jurnal RESPONSIF, vol. 3, no. 1, pp. 37–43, 2021, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti