JIIFKOM (Jurnal Ilmiah Informatika & Komputer) STTR Cepu
E-Mail: iifkom@sttrcepu.ac.id

Analisis Prediksi Risiko Depresi pada Remaja Menggunakan *Random Forest* dan *Logistic Regression* Berdasarkan Pola Penggunaan Media Sosial

Aqid Fahri Hafin[a*], Nilam Tri Astuti[b], Safinatus Zahroh[c]
e-mail: **2408018017@webmail.uad.ac.id**

Magister Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia[a,b]
Magister Teknologi Informasi, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia[c]

Kata kunci:
Depresi, Machine Learning, Random Forest, Logistic Regression, Smote.

Tentang naskah:
-diterima, 17 Juni 2026
-diterbitkan, 1 Juli 2026

Keywords:
Depression, Machine Learning, Random Forest, Logistic Regression, SMOTE.

Intisari

Depresi merupakan gangguan kesehatan mental yang banyak dialami mahasiswa dan dapat memengaruhi kesejahteraan psikologis serta prestasi akademik. Penelitian ini bertujuan mengembangkan dan membandingkan model klasifikasi depresi menggunakan algoritma Logistic Regression dan Random Forest. Dataset yang digunakan terdiri dari 1.200 data dengan distribusi kelas yang tidak seimbang. Tahapan penelitian meliputi pembersihan data, transformasi variabel kategorikal menggunakan *One Hot Encoding*, penyeimbangan data dengan *Synthetic Minority Oversampling Technique* (SMOTE), pembagian data latih dan data uji, serta proses pelatihan dan evaluasi model. Kinerja model diukur menggunakan *accuracy, precision, recall, F1-score*, dan *ROC-AUC*. Hasil pengujian menunjukkan bahwa Random Forest memberikan performa yang lebih baik dibandingkan Logistic Regression dengan *accuracy* 99,79%, *precision* 100%, *recall* 99,57%, *F1-score* 99,79%, dan ROC-AUC 100%. Sementara itu, Logistic Regression memperoleh *accuracy* 96,37%, *precision* 97,38%, *recall* 95,30%, *F1-score* 96,33%, dan ROC-AUC 99,60%. Analisis fitur menunjukkan bahwa durasi tidur, tingkat kecemasan, tingkat stres, dan durasi penggunaan media sosial harian merupakan faktor yang paling berpengaruh terhadap kondisi depresi. Temuan ini menunjukkan bahwa Random Forest efektif digunakan sebagai pendekatan pendukung deteksi dini depresi berdasarkan karakteristik individu dan pola perilaku remaja.

Abstract

Depression is a common mental health disorder among students and can negatively affect psychological well-being and academic performance. This study aims to develop and compare depression classification models using Logistic Regression and Random Forest algorithms. The dataset consisted of 1,200 records with an imbalanced class distribution. The research process included data cleaning, categorical feature transformation using One-Hot Encoding, class balancing through the Synthetic Minority Oversampling Technique (SMOTE), data splitting into training and testing sets, and model training and evaluation. Model performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The results indicate that Random Forest outperformed Logistic Regression, achieving an accuracy of 99.79%, precision of 100%, recall of 99.57%, F1-score of 99.79%, and ROC-AUC of 100%. In comparison, Logistic Regression achieved an accuracy of 96.37%, precision of 97.38%, recall of 95.30%, F1-score of 96.33%, and ROC-AUC of 99.60%. Feature importance analysis revealed that sleep duration, anxiety level, stress level, and daily social media usage were the most influential factors associated with depression. These findings suggest that Random Forest is an effective approach for supporting the early detection of depression based on individual characteristics and behavioral patterns among adolescents.

1. Pendahuluan

Dalam beberapa tahun terakhir, penggunaan media sosial di kalangan remaja mengalami pertumbuhan yang sangat pesat. Secara global, jumlah pengguna aktif media sosial telah mencapai lebih dari 5 miliar pengguna pada tahun 2025 dan terus menunjukkan peningkatan setiap tahunnya (DataReportal, 2025). Kelompok usia remaja merupakan salah satu segmen pengguna terbesar karena tingginya keterlibatan mereka dalam aktivitas digital sehari-hari. Di Indonesia, penggunaan internet dan media sosial juga menunjukkan tren yang signifikan, khususnya pada kelompok usia 15-24 tahun yang merupakan pengguna teknologi digital paling aktif. Berbagai platform media sosial seperti Instagram, TikTok, dan X telah menjadi bagian penting dalam kehidupan remaja sebagai sarana komunikasi, hiburan, interaksi sosial, serta pembentukan identitas diri di lingkungan digital (Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), 2025).

Meskipun media sosial memberikan berbagai manfaat, intensitas penggunaan yang tinggi juga berpotensi menimbulkan dampak negatif terhadap kesehatan mental. Paparan konten yang berlebihan, fenomena perbandingan sosial *(social comparison), cyberbullying*, tekanan sosial digital, serta penggunaan media sosial yang tidak terkendali dapat meningkatkan risiko munculnya berbagai gangguan psikologis, termasuk depresi dan kecemasan. Kondisi ini menjadi perhatian penting mengingat masa remaja merupakan periode perkembangan yang rentan terhadap perubahan emosional dan psikologis.

Deteksi dini depresi pada remaja sering kali menghadapi berbagai kendala karena gejalanya tidak selalu tampak secara langsung dan sulit dikenali melalui observasi konvensional. Selain itu, banyak remaja yang enggan mengungkapkan kondisi psikologisnya akibat stigma sosial maupun rendahnya kesadaran terhadap pentingnya kesehatan mental. Dalam konteks tersebut, pendekatan berbasis data menjadi alternatif yang menjanjikan untuk membantu mengidentifikasi pola risiko depresi secara lebih objektif dan sistematis. Data mengenai kebiasaan penggunaan media sosial, pola tidur, tingkat stres, aktivitas fisik, serta indikator kesehatan mental lainnya dapat dimanfaatkan sebagai dasar dalam membangun sistem prediksi yang mampu mendukung proses deteksi dini.

Perkembangan teknologi *machine learning* membuka peluang yang luas dalam pengembangan model prediksi kesehatan mental. *Machine learning* memungkinkan komputer mempelajari pola dari data historis untuk menghasilkan prediksi terhadap kondisi tertentu secara otomatis (Dong et al., 2025). Dalam bidang kesehatan mental, berbagai algoritma klasifikasi telah digunakan untuk mendeteksi individu yang berisiko mengalami depresi berdasarkan indikator perilaku dan psikologis. Pendekatan ini memiliki keunggulan dalam mengidentifikasi hubungan kompleks antarvariabel yang sering kali sulit ditemukan melalui metode statistik konvensional (Ke et al., 2025).

Meskipun demikian, penelitian terkait prediksi depresi menggunakan *machine learning* masih menunjukkan hasil yang beragam. Perbedaan performa model dipengaruhi oleh karakteristik dataset, jenis fitur yang digunakan, serta algoritma yang diterapkan. Sebagian besar penelitian sebelumnya berfokus pada data klinis, rekam medis, atau hasil survei psikologis, sedangkan penelitian yang memanfaatkan kombinasi indikator perilaku digital dan kesehatan mental pada remaja masih relatif terbatas. Selain itu, studi yang secara khusus membandingkan kinerja Random Forest dan Logistic Regression pada dataset kesehatan mental remaja yang mencakup penggunaan media sosial, pola tidur, aktivitas fisik, tingkat stres, dan kondisi emosional masih belum banyak dilakukan. Oleh karena itu, diperlukan penelitian yang dapat mengevaluasi efektivitas kedua algoritma tersebut dalam memprediksi risiko depresi pada remaja berdasarkan karakteristik perilaku digital dan kondisi kesehatan mental yang lebih komprehensif.

Di antara berbagai algoritma klasifikasi yang tersedia, Random Forest dan Logistic Regression dipilih karena keduanya merepresentasikan pendekatan yang berbeda dalam proses klasifikasi data. Random Forest merupakan metode *ensemble* berbasis decision tree yang mampu menangani hubungan nonlinier, mengurangi risiko overfitting, serta memberikan informasi mengenai tingkat kepentingan fitur *(feature importance)*. Sementara itu, Logistic Regression merupakan model klasifikasi yang sederhana, efisien, dan memiliki tingkat

interpretabilitas yang tinggi sehingga hasil prediksi dapat dijelaskan dengan lebih mudah. Dibandingkan dengan algoritma yang lebih kompleks seperti Support Vector Machine (SVM), XGBoost, maupun Neural Network, kedua metode tersebut menawarkan keseimbangan antara performa prediksi, kebutuhan komputasi yang relatif rendah, dan kemudahan interpretasi yang sangat penting dalam konteks penelitian kesehatan mental (Adefabi et al., 2023).

Berdasarkan permasalahan tersebut, penelitian ini berfokus pada pembangunan model prediksi risiko depresi pada remaja berdasarkan indikator penggunaan media sosial dan kesehatan mental menggunakan *Teenager Mental Health Dataset* yang tersedia secara publik pada platform Kaggle (Algozee, 2024). Algoritma Random Forest dan Logistic Regression diterapkan untuk membangun model klasifikasi, kemudian dibandingkan menggunakan metrik evaluasi *accuracy, precision, recall*, dan *F1-score* sebagaimana digunakan dalam penelitian sebelumnya (Twivy et al., 2025). Kontribusi utama penelitian ini terletak pada evaluasi komparatif kedua algoritma dalam memprediksi risiko depresi pada remaja menggunakan kombinasi indikator perilaku digital dan kesehatan mental. Selain itu, penelitian ini juga bertujuan mengidentifikasi faktor-faktor yang paling berpengaruh terhadap risiko depresi sehingga dapat memberikan pemahaman yang lebih mendalam mengenai hubungan antara perilaku digital dan kondisi psikologis remaja (Afkarinah et al., 2025).

2. Kerangka Teori

2.1. Depresi pada Remaja

Depresi merupakan salah satu gangguan kesehatan mental yang ditandai dengan munculnya perasaan sedih yang berkepanjangan, kehilangan minat terhadap aktivitas yang sebelumnya dianggap menyenangkan, serta penurunan fungsi emosional, kognitif, dan sosial. Kondisi ini tidak hanya memengaruhi suasana hati, tetapi juga dapat berdampak pada pola pikir, perilaku, kualitas tidur, nafsu makan, serta kemampuan individu dalam menjalankan aktivitas sehari-hari. Pada tingkat yang lebih serius, depresi dapat menurunkan produktivitas, memengaruhi hubungan interpersonal, dan meningkatkan risiko perilaku menyakiti diri sendiri (Nagata & others, 2025).

Depresi pada remaja menjadi isu yang perlu mendapat perhatian serius karena periode remaja merupakan tahap perkembangan yang ditandai oleh berbagai perubahan biologis, psikologis, dan sosial yang terjadi secara dinamis. Seperti perubahan hormonal, pencarian identitas diri, tekanan akademik, serta dinamika hubungan dengan keluarga maupun lingkungan sosial dapat meningkatkan kerentanan emosional pada remaja. Ketika kemampuan adaptasi terhadap berbagai perubahan tersebut tidak berjalan dengan baik, risiko munculnya gangguan psikologis, termasuk depresi, menjadi lebih tinggi (Zeng, 2026).

Gejala depresi pada remaja sering kali tidak selalu terlihat secara jelas. Beberapa remaja menunjukkan tanda-tanda seperti mudah marah, menarik diri dari lingkungan sosial, kehilangan motivasi belajar, mengalami gangguan tidur, atau menunjukkan penurunan konsentrasi. Dalam beberapa kasus, gejala depresi juga dapat muncul dalam bentuk kelelahan berkepanjangan, rasa putus asa, serta menurunnya kepercayaan diri. Karena gejalanya dapat menyerupai perubahan perilaku yang umum terjadi pada masa remaja, depresi sering kali terlambat dikenali.

Terdapat berbagai faktor yang dapat meningkatkan risiko depresi pada remaja. Faktor biologis mencakup riwayat keluarga dengan gangguan mental, ketidakseimbangan neurokimia otak, serta perubahan hormonal selama masa pubertas. Faktor psikologis meliputi rendahnya self-esteem, pola pikir negatif, trauma masa lalu, dan kesulitan dalam mengelola emosi. Sementara itu, faktor sosial dapat berupa konflik keluarga, tekanan akademik, isolasi sosial, perundungan, serta kurangnya dukungan dari lingkungan sekitar (Twivy et al., 2025).

Selain faktor-faktor tersebut, perkembangan teknologi digital turut menghadirkan faktor risiko baru. Penggunaan media sosial yang berlebihan dapat memengaruhi kondisi psikologis remaja melalui paparan perbandingan sosial, validasi sosial berbasis interaksi digital, *cyberbullying*, dan tekanan untuk mempertahankan citra diri tertentu. Tingginya intensitas penggunaan media sosial sering kali berhubungan dengan penurunan kualitas tidur serta peningkatan tingkat stres emosional. Oleh karena itu, dalam konteks modern, indikator perilaku digital menjadi salah

satu aspek penting dalam analisis risiko depresi pada remaja (Wu et al., 2025).

Pemahaman mengenai depresi dan faktor-faktor risikonya menjadi landasan penting dalam pengembangan model prediksi berbasis data. Dengan mengidentifikasi indikator yang berkaitan dengan peningkatan risiko depresi, sistem prediksi dapat membantu mendukung proses deteksi dini dan intervensi yang lebih cepat pada remaja yang membutuhkan perhatian khusus.

2.2. Penggunaan Media Sosial

Media sosial merupakan platform digital yang memungkinkan pengguna untuk berinteraksi, berbagi informasi, membangun jejaring sosial, serta mengekspresikan diri secara virtual. Dalam kehidupan remaja modern, media sosial tidak lagi sekadar sarana komunikasi, tetapi telah menjadi bagian dari rutinitas harian yang memengaruhi perilaku, pola pikir, dan dinamika sosial. Platform media sosial seperti Instagram, TikTok, dan X memberikan ruang bagi remaja untuk berinteraksi secara real-time, mengikuti tren, memperoleh hiburan, serta membangun identitas sosial di lingkungan digital (Schønning et al., 2020).

Penggunaan media sosial memiliki dampak yang bersifat dualistik, yaitu dapat memberikan pengaruh positif maupun negatif terhadap kesehatan mental. Dari sisi positif, media sosial dapat membantu remaja memperluas relasi sosial, memperoleh dukungan emosional, meningkatkan rasa keterhubungan sosial, serta menyediakan akses terhadap informasi edukatif. Pada beberapa kasus, media sosial juga menjadi sarana self-expression yang membantu remaja mengekspresikan emosi dan identitas mereka secara lebih terbuka.

Sisi lain dari penggunaan media sosial yang berlebihan dapat meningkatkan berbagai risiko psikologis. Salah satu dampak yang paling sering ditemukan adalah munculnya social comparison, yaitu kecenderungan membandingkan diri dengan kehidupan orang lain yang ditampilkan secara selektif di media sosial. Perbandingan tersebut dapat menurunkan self-esteem dan memunculkan rasa tidak puas terhadap diri sendiri (Khalaf et al., 2023). Selain itu, paparan terhadap komentar negatif, *cyberbullying*, tekanan sosial digital, serta kebutuhan untuk mendapatkan validasi melalui likes atau engagement juga dapat memengaruhi stabilitas emosional remaja.

Penelitian terbaru menunjukkan bahwa hubungan antara media sosial dan kesehatan mental bersifat kompleks dan tidak selalu linear. Beberapa studi menemukan bahwa durasi penggunaan media sosial yang tinggi berkorelasi dengan meningkatnya risiko depresi, kecemasan, dan gangguan tidur. Namun, studi lain menunjukkan bahwa bukan hanya durasi penggunaan yang menjadi faktor utama, melainkan bagaimana media sosial digunakan, jenis interaksi yang terjadi, serta pengalaman emosional pengguna saat berinteraksi di platform digital. Dengan kata lain, kualitas pengalaman digital sering kali lebih menentukan dibandingkan kuantitas waktu penggunaan.

Sejumlah penelitian sebelumnya telah mengeksplorasi keterkaitan antara penggunaan media sosial dan kondisi kesehatan mental pada remaja. Penelitian yang dilakukan oleh Schønning dkk. mengindikasikan adanya hubungan antara aktivitas media sosial dan tingkat kesejahteraan psikologis individu, khususnya terkait depresi, kecemasan, dan *psychological distress*, meskipun kekuatan hubungan tersebut bervariasi antar studi (Schønning et al., 2020). Sementara itu, penelitian Khalaf dkk. menemukan bahwa penggunaan media sosial yang intensif dapat meningkatkan risiko gangguan tidur, kecemasan, dan gejala depresi pada remaja dan dewasa muda (Khalaf et al., 2023).

Penelitian lain oleh Valkenburg dkk. menegaskan bahwa pengaruh media sosial terhadap kesehatan mental tidak dapat digeneralisasi secara sederhana sebagai dampak positif atau negatif. Efek media sosial bergantung pada karakteristik individu, pola penggunaan, tingkat kerentanan psikologis, serta desain platform yang digunakan. Temuan ini menunjukkan bahwa analisis berbasis data menjadi penting untuk memahami indikator mana yang paling berkontribusi terhadap risiko gangguan mental pada remaja (Valkenburg et al., 2022).

Dalam konteks penelitian ini, indikator penggunaan media sosial seperti durasi penggunaan harian, frekuensi interaksi, penggunaan pada malam hari, serta pengalaman negatif selama menggunakan media sosial dipertimbangkan sebagai variabel penting dalam membangun model prediksi depresi. Integrasi indikator

perilaku digital dengan indikator kesehatan mental diharapkan mampu menghasilkan sistem prediksi yang lebih komprehensif dan akurat untuk mendukung deteksi dini risiko depresi pada remaja (Fassi et al., 2024).

2.3. *Machine learning* untuk Klasifikasi

Machine learning merupakan salah satu cabang dari kecerdasan buatan yang memungkinkan sistem komputer untuk mempelajari pola dari data dan menghasilkan prediksi atau keputusan tanpa harus diprogram secara eksplisit untuk setiap skenario. Pendekatan ini banyak digunakan untuk menyelesaikan berbagai permasalahan prediktif, termasuk pada bidang kesehatan, keuangan, pendidikan, hingga analisis perilaku pengguna digital. Kemampuan *machine learning* dalam mengolah data berukuran besar dan kompleks menjadikannya metode yang relevan untuk mendukung pengambilan keputusan berbasis data (Mardini & others, 2025).

Salah satu tugas utama dalam *machine learning* adalah klasifikasi. Klasifikasi merupakan proses pembelajaran yang bertujuan untuk mengelompokkan data ke dalam kategori atau kelas tertentu berdasarkan karakteristik yang dimiliki. Dalam proses ini, model dilatih menggunakan data historis yang telah memiliki label kelas, sehingga model dapat mempelajari hubungan antara variabel input dan output (Do et al., 2025). Setelah proses pelatihan selesai, model digunakan untuk memprediksi kelas dari data baru yang belum pernah dilihat sebelumnya.

Dalam konteks penelitian ini, klasifikasi digunakan untuk memprediksi tingkat risiko depresi pada remaja berdasarkan berbagai indikator penggunaan media sosial dan kondisi kesehatan mental. Masalah ini termasuk dalam klasifikasi biner karena target prediksi terdiri atas dua kelas utama, yaitu remaja dengan risiko depresi dan remaja tanpa risiko depresi (LoParo et al., 2025). Dengan memanfaatkan pendekatan klasifikasi, model dapat membantu mengidentifikasi pola-pola tertentu yang berasosiasi dengan peningkatan risiko depresi.

2.4. *Random Forest*

Random Forest merupakan algoritma *machine learning* yang termasuk dalam metode ensemble learning dan dapat digunakan untuk menyelesaikan permasalahan klasifikasi maupun regresi. Algoritma ini bekerja dengan membangun sejumlah decision tree dari subset data yang berbeda melalui teknik bootstrap sampling. Setiap pohon keputusan akan dilatih menggunakan sebagian data yang dipilih secara acak, kemudian menghasilkan prediksi secara independen. Hasil prediksi dari seluruh pohon akan digabungkan untuk menghasilkan keputusan akhir yang lebih stabil dan akurat.

Keunggulan utama Random Forest dibandingkan single decision tree adalah kemampuannya dalam mengurangi varians model dan meminimalkan risiko overfitting. Selain itu, Random Forest mampu menangani data berdimensi tinggi, data yang mengandung noise, serta memberikan informasi mengenai tingkat kepentingan *(feature importance)* dari setiap variabel yang digunakan dalam proses prediksi (Zhou et al., 2024).

Pada proses pembentukan pohon keputusan, Random Forest menggunakan ukuran impuritas untuk menentukan atribut terbaik sebagai pemisah data. Salah satu ukuran yang umum digunakan adalah *Gini Index* yang dirumuskan pada persamaan (1) berikut:

$$\boldsymbol{Gini(D) = 1 - \sum_{i=1}^{n} p_i^2} \qquad 1$$

Dengan:

$Gini(D)$ = nilai impuritas pada dataset (D)

p_i^2 = probabilitas kemunculan kelas ke-*(i)*

n = jumlah kelas

Semakin kecil nilai Gini Index, maka semakin baik atribut tersebut digunakan sebagai pemisah data pada node pohon keputusan.

Pada kasus klasifikasi, hasil prediksi akhir Random Forest diperoleh melalui mekanisme majority voting, yaitu kelas yang paling banyak dipilih oleh seluruh pohon keputusan akan menjadi hasil klasifikasi akhir (Gohari & others, 2024). majority voting dituliskan sebagai persamaan (2) berikut:

$$\boldsymbol{\hat{Y} = MajorityVote(T_1, T_2, \dots, T_n)} \qquad 2$$

Dengan:

$\hat{Y}$ = hasil prediksi akhir

$T_1, T_2, \ldots, T_n$ = prediksi dari masing-masing pohon keputusan

n = jumlah pohon dalam *Random Forest*

Melalui kombinasi beberapa pohon keputusan dan mekanisme *majority voting*, Random Forest mampu menghasilkan model klasifikasi yang memiliki tingkat akurasi tinggi serta performa yang stabil pada berbagai jenis dataset.

2.5. Logistic Regression

Logistic Regression merupakan salah satu algoritma *supervised learning* yang digunakan untuk memprediksi probabilitas suatu data masuk ke dalam kelas tertentu. Meskipun namanya mengandung kata *regression*, algoritma ini lebih banyak digunakan untuk tugas klasifikasi, khususnya klasifikasi biner. Logistic Regression memodelkan hubungan antara variabel input dan probabilitas output menggunakan fungsi sigmoid, sehingga hasil prediksi berada pada rentang nilai 0 hingga 1 (Zhang et al., 2024).

Model Logistic Regression diawali dengan pembentukan fungsi linear yang menghubungkan seluruh variabel prediktor. Fungsi tersebut dapat dituliskan sebagai persamaan (3) berikut:

$$\boldsymbol{z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n} \qquad 3$$

Dengan:

z = hasil prediksi akhir

β_0 = konstanta *(intercept)*

$\beta_1 \beta_n$ = koefisien variabel

$x_1 x_n$ variabel prediktor

Nilai *(z)* kemudian ditransformasikan menggunakan fungsi *sigmoid* untuk menghasilkan probabilitas suatu data termasuk ke dalam kelas positif. Fungsi *sigmoid* dirumuskan sebagai persamaan (4) berikut:

$$\boldsymbol{P(Y = 1) = \frac{1}{1 + e^{-z}}} \qquad 4$$

Dengan:

$P(Y = 1)$ = probabilitas data termasuk ke kelas positif

e = bilangan eksponensial Euler

z = hasil fungsi linear

Nilai probabilitas yang dihasilkan kemudian dibandingkan dengan nilai ambang batas *(threshold)* tertentu, umumnya sebesar 0,5. Jika nilai probabilitas lebih besar atau sama dengan 0,5 maka data diklasifikasikan ke kelas positif, sedangkan jika nilainya kurang dari 0,5 maka data diklasifikasikan ke kelas negatif.

Keunggulan *Logistic Regression* terletak pada kemampuannya dalam menghasilkan model yang sederhana, mudah diinterpretasikan, serta memiliki waktu komputasi yang relatif cepat. Oleh karena itu, algoritma ini sering digunakan sebagai model dasar (baseline model) dalam penelitian klasifikasi untuk dibandingkan dengan algoritma *machine learning* yang lebih kompleks, seperti Random Forest (Liu & others, 2025).

2.6. Evaluasi Model

Proses evaluasi dalam studi ini bertujuan untuk menilai seberapa baik kedua algoritma yaitu Logistic Regression dan Random Forest dalam menjalankan tugas klasifikasi. Tolok ukur yang digunakan mencakup *Accuracy, Precision, Recall, dan F1-Score*. Seluruh metrik tersebut diturunkan dari elemen-elemen confusion matrix, meliputi *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN), sebagaimana dijelaskan dalam penelitian (Wumaierjiang et al., 2025).

Untuk mengetahui proporsi prediksi yang tepat dibandingkan dengan total keseluruhan data, digunakan metrik *Accuracy* yang perhitungannya mengikuti Persamaan (5).

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}} \qquad 5$$

Adapun *Precision* digunakan untuk mengukur ketepatan model dalam mengklasifikasikan kelas positif, yang dirumuskan melalui Persamaan (6).

$$\boldsymbol{Precision = \frac{TP}{TP + FP}} \qquad 6$$

Sementara itu, *Recall* berfungsi untuk menilai kemampuan model dalam menangkap seluruh data yang benar-benar positif, dengan perhitungan seperti pada Persamaan (7).

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

Tahapan akhir, yaitu *F1-Score* merupakan metrik yang menggabungkan nilai *Precision* dan *Recall* melalui rata-rata harmonik, sehingga dapat digunakan untuk menilai keseimbangan kinerja model dalam kedua aspek tersebut. Perhitungan *F1-Score* dilakukan menggunakan Persamaan (8) berikut.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

3. Metodologi Penelitian

3.1. Dataset Penelitian

Penelitian ini memanfaatkan data sekunder yang bersumber dari *Teenager Mental Health Dataset* yang tersedia pada platform Kaggle (Algozee, 2024). Dataset tersebut berisi data kesehatan mental remaja yang mencakup berbagai variabel, antara lain penggunaan media sosial, durasi tidur, aktivitas fisik, tingkat stres, kondisi emosional, dan kategori risiko depresi. Informasi yang terkandung dalam dataset digunakan untuk mengidentifikasi hubungan antara faktor perilaku dan kondisi psikologis remaja, serta sebagai dasar dalam pengembangan model klasifikasi untuk memprediksi tingkat risiko depresi. Penggunaan dataset ini diharapkan mampu memberikan gambaran yang komprehensif mengenai faktor-faktor yang memengaruhi kesehatan mental remaja dan mendukung proses pembentukan model prediksi yang akurat.

Secara keseluruhan, dataset terdiri atas 1.200 baris observasi yang dilengkapi dengan 13 kolom, di mana 12 di antaranya berperan sebagai variabel masukan (fitur) dan satu sisanya merupakan variabel sasaran (*target*). Adapun variabel target yang digunakan adalah *depression_label*, yang berfungsi untuk menandai tingkat risiko depresi pada tiap individu dalam bentuk dua kelas, yaitu 0 (tidak berisiko depresi) dan 1 (berisiko depresi). Rincian lebih lanjut mengenai komposisi dan karakteristik dataset tersebut disajikan pada Tabel 1.

Tabel 1. Karakteristik Dataset

Keterangan	Nilai
Jumlah Data	1200
Jumlah Fitur	12
Jumlah Kolom Total	13
Target	*depression_label*
Jenis Masalah	Klasifikasi Biner

Dataset ini dipilih karena memiliki variabel yang relevan dengan tujuan penelitian, yaitu menganalisis hubungan antara penggunaan media sosial dan kondisi kesehatan mental remaja terhadap risiko depresi.

3.2. Variabel Dalam Penelitian

Variabel dalam penelitian terdiri dari dua variabel yaitu, variabel independen digunakan sebagai prediktor dalam proses training model *machine learning*, sedangkan variabel dependen merupakan label klasifikasi yang akan diprediksi.

Deskripsi masing-masing variabel ditunjukkan pada Tabel 2.

Tabel 2. Deskripsi Variabel

Variabel	Tipe	Deskripsi
age	Numerik	Usia responden (tahun)
gender	Kategorikal	Jenis kelamin responden
daily_social_media_hours	Numerik	Rata-rata durasi penggunaan media sosial per hari (jam)
platform_usage	Kategorikal	Platform media sosial yang paling sering digunakan
sleep_hours	Numerik	Rata-rata durasi tidur per hari (jam)
screen_time_before_sleep	Numerik	Durasi screen time sebelum tidur (jam)
academic_performance	Numerik	Indikator performa akademik
physical_activity	Numerik	Durasi aktivitas fisik
social_interaction_level	Kategorikal	Tingkat interaksi sosial responden
stress_level	Numerik	Skor tingkat stres
anxiety_level	Numerik	Skor tingkat kecemasan
addiction_level	Numerik	Tingkat kecanduan media sosial
depression_label	Target	Label klasifikasi depresi (0 = Tidak Depresi, 1 = Depresi)

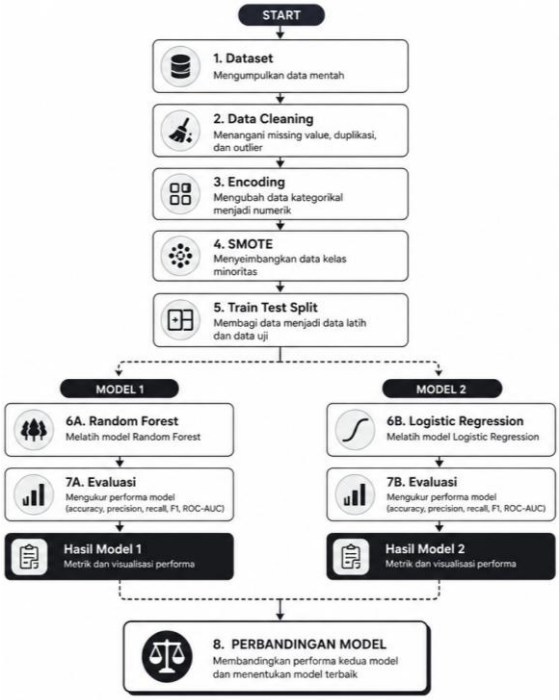
Dalam penelitian ini, variabel seperti tingkat stres, kecemasan, dan kecanduan media sosial diasumsikan memiliki pengaruh signifikan terhadap risiko depresi. Selain itu, variabel perilaku digital seperti durasi penggunaan media sosial dan screen time sebelum tidur juga digunakan sebagai indikator utama dalam proses prediksi.

3.3. Tahapan Penelitian

Untuk membangun dan membandingkan model klasifikasi dalam

memprediksi risiko depresi pada remaja, penelitian ini dirancang melalui prosedur yang terstruktur dengan menguji dua algoritma, yaitu Random Forest dan Logistic Regression. Tahapan diawali dengan pengumpulan dataset yang memuat indikator perilaku digital dan kesehatan mental, kemudian dilanjutkan dengan proses pra-pengolahan data untuk memastikan kualitas informasi yang akan diolah. Setelah data siap, dilakukan pembagian dataset menjadi data latih dan data uji. Kedua algoritma kemudian dilatih menggunakan data latih dan dievaluasi dengan metrik performa yang sama. Hasil evaluasi dari kedua model tersebut selanjutnya dibandingkan untuk menentukan pendekatan mana yang memberikan prediksi paling akurat terhadap tingkat depresi berdasarkan pola penggunaan media sosial dan kondisi psikologis remaja.

Rincian alur metodologi yang diterapkan dalam penelitian ini disajikan secara visual pada Gambar 3.



Gambar 1. Flowchart Tahapan Penelitian

Secara visual, tahapan penelitian dalam penelitian ini mengikuti alur seperti yang disajikan pada Gambar 1. Proses diawali dengan pengumpulan dataset, yang kemudian masuk ke tahap pembersihan data untuk menangani nilai yang hilang, duplikasi, maupun data yang menyimpang. Setelah data bersih, variabel-variabel bertipe kategori dikonversi menjadi representasi numerik melalui proses encoding agar kompatibel dengan algoritma yang digunakan. Untuk mengatasi distribusi kelas yang tidak proporsional, metode SMOTE diterapkan guna menyeimbangkan komposisi data.

Langkah selanjutnya adalah pemisahan dataset menjadi dua bagian, yaitu data latih dan data uji. Data latih berfungsi untuk melatih model, sementara data uji digunakan untuk menilai sejauh mana model mampu bekerja dengan pola data baru yang belum pernah dilihat sebelumnya. Pada titik ini, jalur proses bercabang menjadi dua alur paralel yang masing-masing menggunakan algoritma Random Forest dan Logistic Regression sebagai pendekatan pemodelan.

Setelah proses pelatihan selesai, setiap model dievaluasi menggunakan beberapa metrik klasifikasi standar, seperti akurasi, presisi, recall, F1-score, dan ROC-AUC. Hasil evaluasi dari kedua model tersebut kemudian dibandingkan pada tahap akhir guna menentukan metode mana yang memberikan tingkat akurasi terbaik dalam memprediksi risiko depresi, berdasarkan indikator perilaku dan faktor kesehatan yang diteliti dalam studi ini.

3.4. Preprocessing Data

Tahap pra-pemrosesan data memegang peranan krusial dalam riset *machine learning* guna menyiapkan dataset yang berkualitas sebelum masuk ke fase pelatihan model. Dalam studi ini, serangkaian langkah pra-pemrosesan diterapkan, meliputi proses *encoding, One Hot Encoding*, serta penerapan *Synthetic Minority Oversampling Technique* (SMOTE). Rangkaian proses tersebut bertujuan untuk mengubah data ke dalam format yang dapat diolah, menekan potensi bias akibat distribusi kelas yang timpang, serta mengoptimalkan kapabilitas model dalam menjalankan tugas klasifikasi.

3.4.1. *Encoding*

Encoding merupakan proses transformasi data kategorikal menjadi representasi numerik agar dapat diproses oleh algoritma *machine learning*. Sebagian besar algoritma klasifikasi, termasuk *Random Forest* dan Logistic Regression, tidak dapat mengolah data dalam bentuk teks secara langsung sehingga diperlukan proses konversi ke bentuk numerik.

Pada dataset depresi mahasiswa, beberapa atribut memiliki tipe data kategorikal, seperti jenis kelamin, status pernikahan, jurusan, kebiasaan makan,

dan variabel kategorikal lainnya. Oleh karena itu, dilakukan proses encoding untuk mengubah setiap kategori menjadi representasi numerik yang dapat dikenali oleh model.

3.4.2. *One Hot Encoding*

Metode encoding yang digunakan dalam penelitian ini adalah *One Hot Encoding*. Teknik ini mengubah setiap kategori pada suatu atribut menjadi kolom biner terpisah yang berisi nilai 0 atau 1. Nilai 1 menunjukkan keberadaan kategori tertentu, sedangkan nilai 0 menunjukkan kategori tersebut tidak dipilih.

Sebagai contoh, atribut *Gender* yang memiliki kategori *Male* dan *Female* akan diubah menjadi dua kolom baru, yaitu *Gender_Male* dan *Gender_Female*. Jika suatu data berjenis kelamin laki-laki, maka nilai pada kolom *Gender_Male* bernilai 1 dan *Gender_Female* bernilai 0.

Penggunaan *One Hot Encoding* dipilih karena mampu menghindari munculnya hubungan ordinal semu antar kategori. Jika kategori langsung dikonversi menjadi angka, misalnya laki-laki = 1 dan perempuan = 2, maka model dapat menganggap terdapat hubungan tingkatan antara kedua kategori tersebut. *One Hot Encoding* mengatasi masalah tersebut dengan merepresentasikan setiap kategori secara independen.

3.4.3. SMOTE

Setelah proses *encoding* rampung, langkah selanjutnya adalah mengatasi ketidakseimbangan kelas dengan memanfaatkan metode SMOTE. Keadaan di mana jumlah data pada suatu kelas jauh melampaui kelas lainnya dapat mengakibatkan model lebih condong untuk memprediksi kelas mayoritas dan kurang peka terhadap kelas minoritas.

SMOTE mengatasi hal ini dengan membentuk data sintetis tambahan pada kelas minoritas melalui pendekatan kedekatan antarsampel menggunakan prinsip *k-nearest neighbors* (KNN). Berbeda dengan pendekatan random oversampling yang sekadar menggandakan data lama, SMOTE menciptakan sampel baru yang unik sehingga mampu menekan kemungkinan terjadinya overfitting.

Secara matematis, data sintetis yang dihasilkan oleh SMOTE dapat dihitung menggunakan persamaan (9) berikut.

$$x_{\text{baru}} = x_i + \lambda(x_{nn} - x_i), \quad 0 \le \lambda \le 1 \qquad (9)$$

Dengan:

x_{baru} = sampel sintetis yang dihasilkan

x_i = sampel data minoritas yang dipilih

x_{nn} = salah satu tetangga terdekat (*nearest neighbor*) dari x_i

λ = bilangan acak antara 0 dan 1

Persamaan tersebut menunjukkan bahwa data sintetis dibentuk pada titik di antara sampel minoritas yang dipilih dan salah satu tetangga terdekatnya. Dengan pendekatan ini, SMOTE mampu menambah jumlah data pada kelas minoritas tanpa hanya menduplikasi data yang sudah ada.

Penerapan SMOTE dilakukan setelah proses encoding dan sebelum pembagian data menjadi data latih dan data uji. Tujuan utama penggunaan SMOTE adalah untuk meningkatkan representasi kelas minoritas sehingga model dapat mempelajari pola data secara lebih seimbang.

Distribusi kelas sebelum dan sesudah penerapan SMOTE disajikan pada Tabel 3 dan Tabel 4.

Tabel 3. Distribusi Kelas Sebelum SMOTE

Label	Jumlah
Tidak Depresi	1169
Depresi	31

Tabel 3 menunjukkan distribusi kelas pada dataset sebelum dilakukan proses penyeimbangan data menggunakan metode SMOTE. Terlihat bahwa jumlah data pada kelas Tidak Depresi sebanyak 1.169 data, sedangkan kelas Depresi hanya berjumlah 31 data. Kondisi ini menunjukkan adanya ketidakseimbangan kelas *(class imbalance)* yang sangat signifikan, di mana kelas mayoritas mendominasi sekitar 97,42% dari total data, sedangkan kelas minoritas hanya sekitar 2,58%.

Ketidakseimbangan kelas dapat menyebabkan model *machine learning* cenderung lebih sering memprediksi kelas mayoritas dan mengabaikan pola pada kelas minoritas. Akibatnya, meskipun nilai akurasi terlihat tinggi, kemampuan model dalam mendeteksi kasus depresi dapat menjadi rendah. Oleh karena itu, diperlukan teknik penyeimbangan data untuk meningkatkan representasi kelas minoritas sebelum proses pelatihan model dilakukan.

Tabel 4. Distribusi Kelas Setelah SMOTE

Label	Jumlah
Tidak Depresi	1169
Depresi	1169

Tabel 4 menunjukkan distribusi kelas setelah diterapkan metode SMOTE. Metode ini bekerja dengan menghasilkan sampel sintetis pada kelas minoritas berdasarkan karakteristik data yang sudah ada, sehingga jumlah data pada kedua kelas menjadi seimbang.

Setelah proses SMOTE dilakukan, jumlah data pada kelas Depresi meningkat dari 31 menjadi 1.169 data, sehingga sama dengan jumlah data pada kelas Tidak Depresi. Dengan demikian, distribusi kelas menjadi 1:1 atau seimbang. Kondisi ini diharapkan dapat membantu model *machine learning* mempelajari karakteristik kedua kelas secara lebih baik dan mengurangi bias terhadap kelas mayoritas.

Penerapan SMOTE penting dalam penelitian ini karena dapat meningkatkan kemampuan model dalam mengidentifikasi individu yang mengalami depresi. Selain itu, data yang seimbang memungkinkan proses pelatihan *Random Forest* dan Logistic Regression menghasilkan model yang lebih adil dan memiliki performa klasifikasi yang lebih baik, terutama pada metrik *recall*, *precision*, dan F1-score untuk kelas minoritas.

3.5. Implementasi *Random Forest*

Pada penelitian ini, model *Random Forest* dibangun menggunakan beberapa parameter yang disesuaikan untuk memperoleh performa klasifikasi yang optimal. Parameter yang digunakan dalam proses pelatihan model ditunjukkan pada Tabel 5.

Tabel 5. Parameter *Random Forest*

Parameter	Nilai	Deskripsi
n_estimators	200	Menunjukkan jumlah pohon keputusan (*decision tree*) yang dibangun dalam model *Random Forest* sebanyak 200 pohon.
max_depth	None	Tidak terdapat batasan kedalaman pohon sehingga setiap pohon dapat berkembang hingga memenuhi kriteria penghentian.
min_samples_split	2	Menentukan jumlah minimum sampel yang diperlukan untuk melakukan pemisahan (*split*) pada suatu node.
min_samples_leaf	1	Menentukan jumlah minimum sampel yang harus terdapat pada node daun (*leaf node*).
max_features	sqrt	Menentukan jumlah fitur yang dipilih secara acak pada setiap proses pembentukan node sebesar akar kuadrat dari total fitur yang tersedia.

Konfigurasi parameter tersebut digunakan untuk menghasilkan model *Random Forest* yang mampu mempelajari pola data secara efektif sekaligus menjaga kemampuan generalisasi terhadap data yang belum pernah dilihat sebelumnya.

3.6. Implementasi Logistic Regression

Logistic Regression digunakan sebagai model pembanding terhadap *Random Forest* untuk mengetahui algoritma yang memberikan performa terbaik dalam klasifikasi tingkat depresi mahasiswa. Parameter yang digunakan dalam proses pelatihan model ditunjukkan pada Tabel 6.

Tabel 6. Parameter Logistic Regression

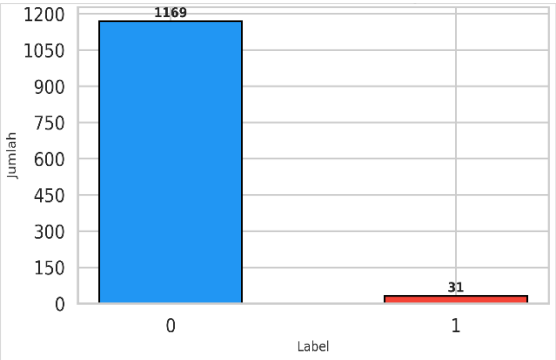
Parameter	Nilai	Deskripsi
penalty	l2	Digunakan untuk menerapkan regularisasi L2 guna mengurangi risiko *overfitting* pada model.
C	1.0	Merupakan nilai invers dari kekuatan regularisasi. Semakin kecil nilai C, semakin kuat regularisasi yang diterapkan.
solver	lbfgs	Algoritma optimasi yang digunakan untuk menemukan parameter model yang optimal selama proses pelatihan.
max_iter	1000	Menunjukkan jumlah iterasi maksimum yang diperbolehkan selama proses optimasi model.
class_weight	None	Tidak memberikan bobot khusus pada kelas tertentu karena ketidakseimbangan data telah ditangani menggunakan metode SMOTE.

Parameter tersebut dipilih untuk menghasilkan model Logistic Regression yang stabil dan mampu melakukan konvergensi dengan baik selama proses pelatihan. Selanjutnya, performa kedua model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, F1-score, dan ROC-AUC, kemudian dibandingkan untuk menentukan model terbaik dalam klasifikasi depresi mahasiswa.

4. Hasil dan Pembahasan

4.1. Analisis Dataset

Analisis dataset dilakukan untuk memahami karakteristik data sebelum model *machine learning* dibangun. Tahap ini bertujuan untuk mengidentifikasi distribusi label target, pola hubungan antar variabel, serta potensi permasalahan pada dataset yang dapat memengaruhi performa model. Pada penelitian ini, analisis dilakukan menggunakan visualisasi distribusi label depresi dan heatmap korelasi antar fitur numerik.

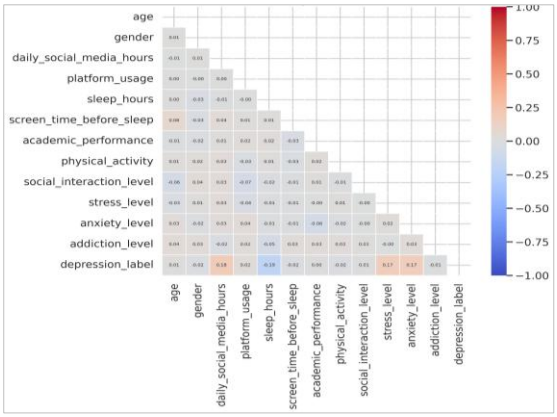


Gambar 2. Distribusi Label Depresi

Gambar 2 menunjukkan distribusi label pada variabel target *depression_label*. Label 0 merepresentasikan remaja yang tidak berisiko depresi, sedangkan label 1 merepresentasikan remaja yang berisiko depresi.

Berdasarkan *bar chart* pada Gambar 2, terlihat bahwa dataset memiliki distribusi kelas yang sangat tidak seimbang *(imbalanced dataset)*. Jumlah data pada kelas 0 (tidak depresi) sebanyak 1169 data, sedangkan kelas 1 (depresi) hanya sebanyak 31 data. Hal ini menunjukkan bahwa sekitar 97,42% data berada pada kelas non-depresi, sementara hanya 2,58% yang termasuk kelas depresi.

Ketidakseimbangan distribusi kelas ini dapat menjadi tantangan dalam proses klasifikasi karena model cenderung bias terhadap kelas mayoritas. Dalam kondisi seperti ini, model berpotensi menghasilkan nilai *accuracy* yang tinggi tetapi gagal mendeteksi kelas minoritas dengan baik. Oleh karena itu, metrik evaluasi seperti *precision*, *recall*, F1-score, dan ROC-AUC menjadi penting untuk memberikan penilaian performa model yang lebih komprehensif dibanding hanya menggunakan *accuracy*.



Gambar 3. Heatmap Korelasi

Gambar 3 menunjukkan heatmap korelasi antar fitur numerik dalam dataset. Visualisasi ini digunakan untuk melihat kekuatan hubungan linear antar variabel menggunakan koefisien korelasi Pearson dengan rentang nilai dari (-1) hingga 1. Nilai mendekati 1 menunjukkan korelasi positif kuat, nilai mendekati (-1) menunjukkan korelasi negatif kuat, sedangkan nilai mendekati 0 menunjukkan hubungan linear yang lemah.

Berdasarkan heatmap, sebagian besar fitur numerik memiliki korelasi yang relatif rendah satu sama lain, yang ditunjukkan oleh nilai korelasi yang mendekati nol. Hal ini mengindikasikan bahwa tidak terdapat multikolinearitas yang kuat antar fitur, sehingga sebagian besar variabel dapat digunakan secara bersamaan dalam proses pelatihan model tanpa menyebabkan redundansi informasi yang signifikan.

Jika dilihat terhadap variabel target depression_label, beberapa fitur menunjukkan korelasi yang relatif lebih menonjol dibanding fitur lainnya. Variabel *daily_social_media_hours* memiliki korelasi positif sebesar 0,18, yang menunjukkan bahwa semakin tinggi durasi penggunaan media sosial harian, maka kecenderungan risiko depresi juga meningkat. Variabel *stress_level* dan *anxiety_level* juga menunjukkan korelasi positif masing-masing sebesar 0,17, yang mengindikasikan bahwa peningkatan

tingkat stres dan kecemasan berkaitan dengan meningkatnya risiko depresi.

Sebaliknya, variabel *sleep_hours* memiliki korelasi negatif sebesar -0,19 terhadap label depresi. Hal ini menunjukkan bahwa semakin sedikit durasi tidur remaja, maka risiko depresi cenderung meningkat. Temuan ini konsisten dengan literatur kesehatan mental yang menyebutkan bahwa kualitas dan durasi tidur memiliki hubungan erat dengan stabilitas emosional.

Secara keseluruhan, analisis korelasi menunjukkan bahwa meskipun tidak terdapat hubungan linear yang sangat kuat, fitur-fitur seperti durasi penggunaan media sosial, tingkat stres, tingkat kecemasan, dan durasi tidur memiliki potensi kontribusi penting dalam proses prediksi risiko depresi pada remaja. variabel-variabel tersebut selanjutnya akan digunakan dalam pembangunan model klasifikasi menggunakan algoritma Random Forest dan Logistic Regression

4.2. Hasil Evaluasi *Random Forest*

Setelah melalui tahap preprocessing, penyeimbangan data menggunakan SMOTE, dan proses pelatihan model, algoritma Random Forest diuji menggunakan data pengujian untuk mengetahui kemampuannya dalam memprediksi risiko depresi pada remaja. Evaluasi model dilakukan menggunakan beberapa metrik, yaitu *Accuracy*, *Precision*, *Recall*, F1-Score, dan ROC-AUC. Hasil evaluasi kinerja Random Forest ditunjukkan pada Tabel 7.

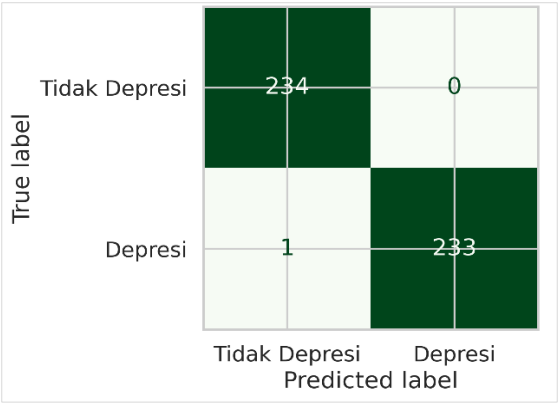
Tabel 7. Hasil Evaluasi Random Forest

Metrik	Nilai
Accuracy	99,79%
Precision	100,00%
Recall	99,57%
F1-Score	99,79%
ROC-AUC	100,00%

Berdasarkan hasil pengujian yang ditunjukkan pada Tabel 7, algoritma Random Forest memperoleh nilai *Accuracy* sebesar 99,79%, *Precision* sebesar 100%, *Recall* sebesar 99,57%, F1-Score sebesar 99,79%, dan ROC-AUC sebesar 100%. Hasil tersebut menunjukkan bahwa model Random Forest mampu melakukan klasifikasi risiko depresi pada remaja dengan tingkat akurasi yang sangat tinggi berdasarkan indikator penggunaan media sosial dan kesehatan mental.

Nilai *Precision* sebesar 100% menunjukkan bahwa seluruh data yang diprediksi sebagai kategori depresi merupakan data yang benar-benar termasuk dalam kategori depresi. Sementara itu, nilai *Recall* sebesar 99,57% menunjukkan bahwa hampir seluruh kasus depresi berhasil diidentifikasi oleh model. Tingginya nilai F1-Score mengindikasikan bahwa model memiliki keseimbangan yang baik antara *Precision* dan *Recall* sehingga mampu memberikan hasil klasifikasi yang konsisten.

Untuk melihat distribusi hasil prediksi yang dihasilkan oleh model Random Forest, digunakan Confusion Matrix sebagaimana ditunjukkan pada Gambar 4.



Gambar 4. Confusion Matrix Random Forest

Berdasarkan Gambar 4, model berhasil mengklasifikasikan 234 data kategori Tidak Depresi dan 233 data kategori Depresi dengan benar. Selain itu, tidak ditemukan kesalahan klasifikasi pada kategori Tidak Depresi (*False Positive* = 0). Namun, terdapat satu data kategori Depresi yang diprediksi sebagai Tidak Depresi (*False Negative* = 1). Hasil ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan remaja yang berisiko mengalami depresi dan yang tidak mengalami depresi.

Performa yang tinggi pada algoritma Random Forest menunjukkan bahwa kombinasi variabel seperti daily social media hours, sleep hours, stress level, anxiety level, dan addiction level memiliki pola yang dapat dipelajari dengan baik oleh model. Dengan demikian, Random Forest dapat digunakan sebagai salah satu pendekatan yang efektif dalam mendukung deteksi dini risiko depresi pada remaja berdasarkan aktivitas

penggunaan media sosial dan kondisi kesehatan mental.

4.3. Hasil Evaluasi Logistic Regression

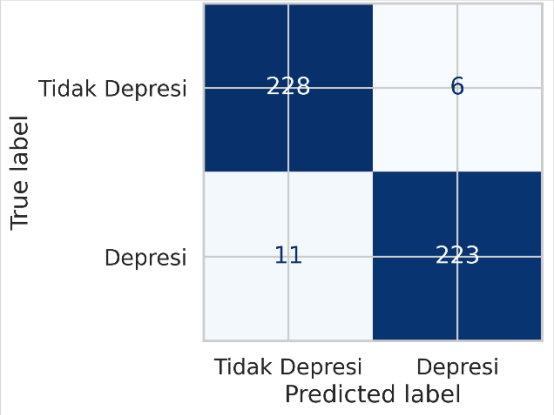
Setelah dilakukan proses pelatihan dan pengujian menggunakan algoritma Logistic Regression, model dievaluasi untuk mengetahui kemampuannya dalam mengklasifikasikan risiko depresi pada remaja berdasarkan indikator penggunaan media sosial dan kesehatan mental. Evaluasi dilakukan menggunakan metrik *Accuracy*, *Precision*, *Recall*, F1-Score, dan ROC-AUC. Hasil evaluasi model Logistic Regression ditunjukkan pada Tabel 8.

Tabel 8. Hasil Evaluasi Logistic Regression

Metrik	Nilai
Accuracy	96,41%
Precision	97,44%
Recall	95,40%
F1-Score	96,41%
ROC-AUC	96,41%

Berdasarkan hasil pengujian pada Tabel 8, algoritma Logistic Regression menunjukkan performa yang sangat baik dalam melakukan klasifikasi risiko depresi. Nilai *Accuracy* sebesar 96,41% menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar oleh model. Selain itu, nilai *Precision* sebesar 97,44% menunjukkan bahwa mayoritas data yang diprediksi sebagai depresi merupakan data yang benar-benar termasuk kategori depresi. Nilai *Recall* sebesar 95,40% mengindikasikan bahwa sebagian besar kasus depresi berhasil teridentifikasi oleh model.

Untuk melihat distribusi hasil prediksi secara lebih rinci, digunakan Confusion Matrix sebagaimana ditunjukkan pada Gambar 5.



Gambar 5. Confusion Matrix Logistic Regression

Berdasarkan Gambar 5, model Logistic Regression mampu mengklasifikasikan dengan benar sebanyak 228 data pada kategori Tidak Depresi dan 228 data pada kategori Depresi. Meskipun demikian, masih ditemukan kesalahan klasifikasi berupa 6 data kategori Tidak Depresi yang diprediksi sebagai Depresi *(False Positive)* serta 11 data kategori Depresi yang diprediksi sebagai Tidak Depresi (*False Negative)*. Temuan ini menunjukkan bahwa terdapat sejumlah pola data yang belum dapat dipisahkan secara optimal oleh model Logistic Regression.

Secara keseluruhan, hasil evaluasi mengindikasikan bahwa Logistic Regression memiliki kemampuan klasifikasi yang baik dengan tingkat kesalahan yang relatif rendah. Tingginya jumlah prediksi yang benar menunjukkan bahwa model mampu mempelajari hubungan antara variabel penggunaan media sosial, kondisi psikologis, dan risiko depresi secara efektif. Namun, karena Logistic Regression merupakan model linear, kemampuannya dalam menangkap interaksi dan hubungan nonlinier yang kompleks antarvariabel masih terbatas. Oleh karena itu, performanya cenderung lebih rendah dibandingkan metode berbasis *ensemble* seperti Random Forest yang mampu mengakomodasi pola data yang lebih kompleks dan beragam.

4.4. Perbandingan Kinerja Model

Setelah dilakukan proses pelatihan dan evaluasi menggunakan algoritma Logistic Regression dan Random Forest, diperoleh hasil kinerja model yang disajikan pada Tabel 9.

Tabel 9. Perbandingan Kinerja Model

Model	*Logistic Regression*	*Random Forest*
Accuracy	0,9637	0,9979
Precision	0,9738	1,0000
Recall	0,9530	0,9957
F1-Score	0,9633	0,9979
OC-AUC	0,9960	1,0000

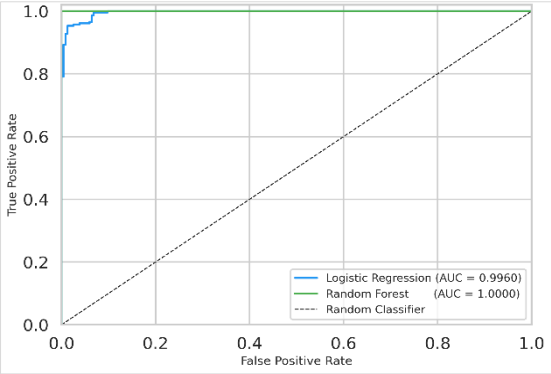
Berdasarkan Tabel 9, kedua model menunjukkan performa klasifikasi yang sangat baik dalam mendeteksi tingkat depresi. Namun, algoritma Random Forest menghasilkan nilai yang lebih tinggi pada seluruh metrik evaluasi dibandingkan Logistic Regression.

Model Logistic Regression memperoleh nilai akurasi sebesar 96,37%, *precision*

sebesar 97,38%, *recall* sebesar 95,30%, F1-score sebesar 96,33%, dan ROC-AUC sebesar 99,60%. Hasil ini menunjukkan bahwa Logistic Regression mampu mengklasifikasikan data dengan baik meskipun menggunakan pendekatan linier yang relatif sederhana.

Sementara itu, Random Forest memperoleh akurasi sebesar 99,79%, *precision* sebesar 100%, *recall* sebesar 99,57%, F1-score sebesar 99,79%, dan ROC-AUC sebesar 100%. Nilai tersebut menunjukkan bahwa hampir seluruh data berhasil diklasifikasikan dengan benar. Tingginya nilai *precision* menunjukkan bahwa model mampu meminimalkan kesalahan prediksi positif, sedangkan nilai *recall* yang sangat tinggi menunjukkan kemampuan model dalam mendeteksi kasus depresi secara efektif.

Perbedaan performa ini menunjukkan bahwa Random Forest lebih mampu menangkap hubungan kompleks antar variabel dibandingkan Logistic Regression. Dengan memanfaatkan mekanisme *ensemble learning* melalui banyak pohon keputusan, Random Forest dapat menghasilkan model yang lebih robust dan memiliki kemampuan generalisasi yang lebih baik pada dataset yang digunakan dalam penelitian ini.



Gambar 6. ROC Curve Logistic Reggression vc Random Fores

Berdasarkan Gambar 6, kurva ROC Random Forest berada lebih dekat ke sudut kiri atas dibandingkan Logistic Regression. Hal ini menunjukkan bahwa Random Forest memiliki kemampuan yang lebih baik dalam membedakan kelas depresi dan tidak depresi pada berbagai nilai ambang klasifikasi (threshold).

Nilai ROC-AUC Random Forest sebesar 1,000 mengindikasikan kemampuan diskriminasi yang hampir sempurna, sedangkan Logistic Regression memperoleh nilai ROC-AUC sebesar 0,996 yang juga termasuk kategori sangat baik. Meskipun perbedaannya relatif kecil, hasil ini memperkuat temuan pada Tabel 9 bahwa Random Forest merupakan model terbaik pada penelitian ini.

4.5. Analisis Faktor yang Berpengaruh

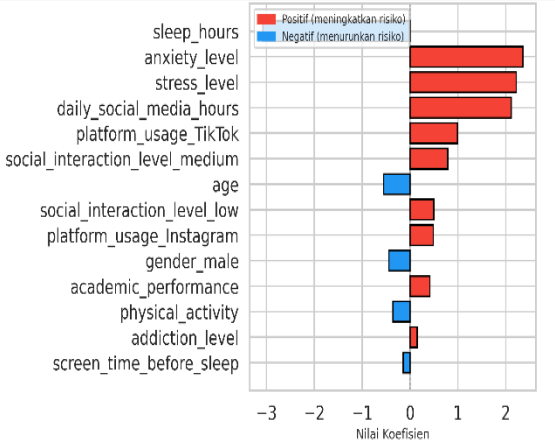
Tidak hanya berfokus pada performa klasifikasi, penelitian ini juga mengkaji faktor-faktor yang berperan dalam memengaruhi kondisi depresi. Analisis dilakukan dengan memanfaatkan informasi koefisien pada Logistic Regression dan tingkat kepentingan fitur (*feature importance*) pada Random Forest untuk mengidentifikasi variabel yang paling berkontribusi terhadap hasil prediksi. Pada Logistic Regression, tingkat pengaruh variabel dianalisis menggunakan nilai koefisien absolut (absolute coefficient), sedangkan pada Random Forest digunakan nilai feature importance.

Tabel 10. Perbandingan Pengaruh Variabel pada Logistic Regression dan Random Forest

Variabel	Koef_LR	AbsKoef_LR	Importance_RF
sleep_hours	-3.0761	3.0761	0.2595
anxiety_level	2.3589	2.3589	0.2021
stress_level	2.2199	2.2199	0.2382
daily_social_media_hours	2.1148	2.1148	0.1715
platform_usage_TikTok	0.9922	0.9922	0.0353
social_interaction_level_medium	0.7854	0.7854	0.0036
age	-0.5495	0.5495	0.0054
social_interaction_level_low	0.4956	0.4956	0.0037
platform_usage_Instagram	0.4851	0.4851	0.0067
gender_male	-0.4433	0.4433	0.0018
academic_performance	0.4156	0.4156	0.0273
physical_activity	-0.3600	0.3600	0.0196
addiction_level	0.1492	0.1492	0.0093
screen_time_before_sleep	-0.1457	0.1457	0.0159

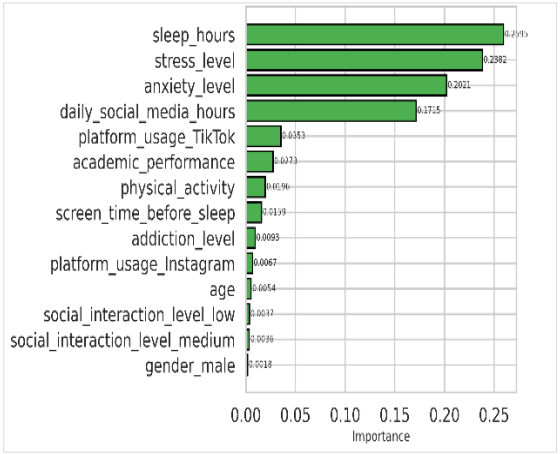
Berdasarkan Tabel 10, variabel yang memiliki pengaruh terbesar pada model Logistic Regression adalah *sleep_hours, anxiety_level, stress_level,* dan *daily_social_media_hours.* Nilai koefisien negatif pada variabel *sleep_hours* menunjukkan bahwa semakin lama waktu tidur seseorang, semakin rendah kemungkinan mengalami depresi. Sebaliknya, variabel *anxiety_level, stress_level, dan daily_social_media_hours* memiliki koefisien positif yang menunjukkan bahwa peningkatan nilai pada variabel tersebut cenderung meningkatkan risiko depresi.

Hasil yang diperoleh dari Random Forest menunjukkan pola yang serupa. Variabel *sleep_hours* memiliki nilai feature importance tertinggi sebesar 0,2595, diikuti oleh *stress_level* sebesar 0,2382, *anxiety_level* sebesar 0,2021, dan *daily_social_media_hours* sebesar 0,1715. Keempat variabel tersebut secara konsisten muncul sebagai faktor dominan pada kedua model, sehingga dapat dianggap sebagai faktor utama yang memengaruhi kondisi depresi pada dataset penelitian.



Gambar 7. Feature Importance Logistic Regression

Berdasarkan Gambar 7, variabel *sleep_hours* memiliki nilai koefisien absolut tertinggi, yang menunjukkan bahwa durasi tidur merupakan faktor paling berpengaruh dalam prediksi depresi. Selain itu, *variabel anxiety_level, stress_level,* dan *daily_social_media_hours* juga memiliki kontribusi yang besar terhadap hasil klasifikasi. Temuan ini menunjukkan bahwa faktor psikologis dan pola penggunaan media sosial memiliki hubungan yang kuat dengan kondisi depresi.



Gambar 8. Feature Importance Random Forest

Gambar 8 menunjukkan variabel *sleep_hours* menempati posisi pertama sebagai faktor yang paling berpengaruh terhadap prediksi depresi, diikuti oleh *stress_level, anxiety_level,* dan *daily_social_media_hours.* Hasil ini memperkuat temuan dari Logistic Regression bahwa kualitas pola tidur, tingkat stres, tingkat kecemasan, dan intensitas penggunaan media sosial merupakan faktor dominan yang berkaitan dengan kondisi depresi.

Secara keseluruhan, kedua model menghasilkan pola interpretasi yang relatif konsisten. Kesamaan hasil tersebut meningkatkan kepercayaan terhadap temuan penelitian bahwa faktor psikologis dan gaya hidup digital memiliki peran penting dalam memengaruhi tingkat depresi. Dengan demikian, upaya pencegahan depresi dapat difokuskan pada pengelolaan stres dan kecemasan, peningkatan kualitas tidur, serta penggunaan media sosial yang lebih sehat dan terkontrol.

4.6. Pembahasan

4.6.1 Analisis Performa Model

Penelitian ini bertujuan untuk membangun model klasifikasi depresi menggunakan algoritma Logistic Regression dan Random Forest, serta mengidentifikasi faktor-faktor yang paling berpengaruh terhadap kondisi depresi. Sebelum proses pemodelan dilakukan, dataset terlebih dahulu melalui tahapan preprocessing yang meliputi data *cleaning, One Hot Encoding,* dan penanganan ketidakseimbangan kelas menggunakan metode SMOTE. Penerapan SMOTE berhasil menyeimbangkan distribusi kelas

dari 1.169 data tidak depresi dan 31 data depresi menjadi masing-masing 1.169 data, sehingga model dapat mempelajari karakteristik kedua kelas secara lebih optimal.

Hasil evaluasi menunjukkan bahwa kedua algoritma memiliki performa yang sangat baik dalam melakukan klasifikasi depresi. Logistic Regression memperoleh nilai *accuracy* sebesar 96,37%, *precision* sebesar 97,38%, *recall* sebesar 95,30%, F1-score sebesar 96,33%, dan ROC-AUC sebesar 99,60%. Sementara itu, Random Forest menghasilkan performa yang lebih tinggi dengan *accuracy* sebesar 99,79%, *precision* sebesar 100%, *recall* sebesar 99,57%, F1-score sebesar 99,79%, dan ROC-AUC sebesar 100%.

Keunggulan Random Forest dibandingkan Logistic Regression menunjukkan bahwa hubungan antara variabel-variabel pada dataset depresi cenderung bersifat kompleks dan tidak sepenuhnya linier. Sebagai metode *ensemble learning*, Random Forest mampu membangun banyak pohon keputusan dan menggabungkan hasil prediksinya sehingga lebih efektif dalam menangkap pola data yang beragam. Hal ini terlihat dari peningkatan nilai *recall* dan F1-score yang menunjukkan kemampuan model dalam mendeteksi kasus depresi secara lebih akurat.

Analisis faktor yang berpengaruh menunjukkan bahwa durasi tidur *(sleep_hours)*, tingkat kecemasan *(anxiety_level),* tingkat stres *(stress_level)*, dan durasi penggunaan media sosial harian *(daily_social_media_hours*) merupakan variabel yang paling dominan pada kedua model. Variabel (*sleep_hours*) memiliki pengaruh terbesar dengan koefisien negatif pada Logistic Regression dan nilai feature importance tertinggi pada Random Forest. Hasil ini menunjukkan bahwa semakin tinggi durasi tidur, semakin rendah kemungkinan seseorang mengalami depresi. Sebaliknya, peningkatan tingkat kecemasan, tingkat stres, dan penggunaan media sosial harian cenderung meningkatkan risiko depresi.

Temuan tersebut sejalan dengan berbagai penelitian sebelumnya yang menunjukkan bahwa gangguan pola tidur, stres, dan kecemasan merupakan faktor utama yang berkaitan dengan kesehatan mental. Selain itu, intensitas penggunaan media sosial yang tinggi juga dapat meningkatkan risiko depresi akibat paparan informasi yang berlebihan, perbandingan sosial, maupun berkurangnya interaksi sosial secara langsung.

Berdasarkan seluruh hasil yang diperoleh, dapat disimpulkan bahwa Random Forest merupakan model terbaik untuk klasifikasi depresi pada penelitian ini karena memberikan performa evaluasi yang lebih tinggi dibandingkan Logistic Regression. Selain itu, hasil analisis faktor menunjukkan bahwa aspek psikologis, kualitas tidur, dan perilaku penggunaan media sosial memiliki peran penting dalam memengaruhi kondisi depresi. Temuan ini diharapkan dapat menjadi dasar dalam pengembangan sistem deteksi dini depresi serta penyusunan strategi pencegahan yang lebih efektif bagi kelompok berisiko.

4.6.1 Analisis Potensi *Overfitting*

Nilai akurasi sebesar 99,79% dan ROC-AUC sebesar 100% menunjukkan kemampuan klasifikasi yang sangat tinggi. Meskipun demikian, hasil tersebut perlu diinterpretasikan secara hati-hati karena performa yang sangat tinggi dapat mengindikasikan potensi *overfitting*. Dalam penelitian ini, evaluasi model dilakukan menggunakan data uji yang terpisah dari data pelatihan sehingga model tidak dievaluasi menggunakan data yang sama dengan data pelatihan.

Selain itu, Random Forest memiliki mekanisme *bootstrap aggregation (bagging)* yang dapat membantu mengurangi risiko *overfitting* dibandingkan decision tree tunggal. Tingginya performa model juga dapat dipengaruhi oleh karakteristik dataset yang memiliki pemisahan kelas yang cukup jelas berdasarkan kombinasi indikator kesehatan mental dan perilaku digital. Namun demikian, penelitian selanjutnya disarankan untuk menerapkan *k-fold cross validation* atau menggunakan dataset eksternal guna menguji kemampuan generalisasi model secara lebih komprehensif.

4.6.2 Perbandingan dengan Penelitian Terdahulu

Hasil penelitian ini menunjukkan bahwa Random Forest memberikan performa yang lebih baik dibandingkan Logistic Regression dalam klasifikasi risiko depresi. Temuan ini sejalan dengan penelitian (Adefabi et al., 2023) yang menunjukkan bahwa Random Forest mampu menghasilkan performa klasifikasi yang lebih tinggi dibandingkan model linear karena kemampuannya dalam

menangkap hubungan nonlinier antarvariabel. Hasil serupa juga dilaporkan oleh (Twivy et al., 2025), yang menemukan bahwa metode berbasis *ensemble* memiliki tingkat akurasi yang lebih baik dalam prediksi kondisi kesehatan mental dibandingkan pendekatan klasifikasi konvensional. Dibandingkan dengan penelitian-penelitian tersebut, model yang dikembangkan pada penelitian ini menghasilkan tingkat akurasi yang relatif lebih tinggi. Perbedaan tersebut dapat dipengaruhi oleh karakteristik dataset, kualitas fitur yang digunakan, proses preprocessing data, serta penerapan metode SMOTE untuk mengatasi ketidakseimbangan kelas.

4. Kesimpulan

Penelitian ini berhasil membangun model klasifikasi depresi menggunakan algoritma Logistic Regression dan Random Forest pada dataset kesehatan mental remaja. Tahapan preprocessing yang meliputi pembersihan data, transformasi fitur menggunakan *One Hot Encoding*, serta penanganan ketidakseimbangan kelas dengan metode SMOTE terbukti mampu menghasilkan dataset yang lebih representatif untuk proses pelatihan model.

Berdasarkan hasil evaluasi, kedua algoritma menunjukkan performa yang sangat baik dalam mengklasifikasikan kondisi depresi. Namun demikian, Random Forest memberikan kinerja yang lebih unggul dibandingkan Logistic Regression pada seluruh metrik evaluasi, sehingga dapat dianggap sebagai model yang paling efektif untuk kasus klasifikasi depresi pada dataset yang digunakan. Hasil penelitian juga menunjukkan bahwa durasi tidur *(sleep_hours),* tingkat kecemasan *(anxiety_level),* tingkat stres *(stress_level),* dan durasi penggunaan media sosial harian *(daily_social_media_hours)* merupakan faktor-faktor yang paling berpengaruh terhadap risiko depresi.

Temuan ini menunjukkan bahwa kombinasi indikator kesehatan mental, kualitas tidur, dan perilaku penggunaan media sosial dapat dimanfaatkan secara efektif dalam membangun sistem prediksi depresi berbasis machine learning. Dengan demikian, model yang dihasilkan berpotensi mendukung upaya deteksi dini depresi pada remaja sehingga intervensi preventif dapat dilakukan secara lebih cepat dan tepat sasaran.

5. Saran

Terdapat sejumlah kendala dalam studi ini yang sekaligus membuka ruang perbaikan untuk riset mendatang. Pengayaan jumlah sampel serta variasi latar belakang responden pada dataset sangat diperlukan agar model yang dikembangkan memiliki kapasitas generalisasi yang lebih baik terhadap populasi secara keseluruhan.

Penelitian selanjutnya disarankan untuk menggunakan dataset dengan jumlah sampel yang lebih besar dan karakteristik responden yang lebih beragam guna meningkatkan kemampuan generalisasi model. Selain itu, validasi model dapat diperkuat melalui penerapan teknik *K-Fold Cross Validation* maupun pengujian menggunakan dataset eksternal untuk memastikan performa model pada data yang berbeda.

Pengembangan penelitian juga dapat dilakukan dengan membandingkan algoritma lain seperti XGBoost, LightGBM, Support Vector Machine (SVM), maupun *Artificial Neural Network* (ANN). Optimalisasi model melalui hyperparameter tuning *menggunakan Grid Search* atau *Random Search* juga berpotensi meningkatkan performa klasifikasi.

Selain faktor yang telah digunakan dalam penelitian ini, studi selanjutnya dapat mempertimbangkan variabel tambahan seperti kondisi sosial ekonomi, lingkungan keluarga, aktivitas akademik, serta riwayat kesehatan mental untuk memperoleh pemahaman yang lebih komprehensif mengenai faktor-faktor yang memengaruhi depresi. Hasil penelitian ini juga dapat dikembangkan menjadi sistem atau aplikasi deteksi dini depresi berbasis machine learning yang dapat digunakan sebagai alat bantu dalam mendukung pemantauan kesehatan mental remaja.

Daftar Pustaka

Adefabi, A., Olisah, S., Obunadike, C., Oyetubo, O., Taiwo, E., & Tella, E. (2023). Predicting Accident Severity: An Analysis of Factors Affecting Accident Severity Using Random Forest Model. *International Journal on Cybernetics & Informatics*, *12*(6), 107–121. https://doi.org/10.5121/ijci.2023.120609

Afkarinah, I., Faizin, A., & Havy, A. Z. F. (2025). *Klasifikasi Tingkat Keparahan Kecelakaan Kerja Menggunakan Algoritma Random Forest*. https://doi.org/https://doi.org/10.61722/jssr.v3i6.6433

Algozee. (2024). *Teenager Mental Health Dataset*. Kaggle. https://www.kaggle.com/datasets/algozee/teenager-menthal-healy/data

Asosiasi Penyelenggara Jasa Internet Indonesia (APJII). (2025). *Survei Profil Internet Indonesia 2025*. https://apjii.or.id

DataReportal. (2025). *Digital 2025 Indonesia*. https://datareportal.com/reports/digital-2025-indonesia

Do, H., Huang, K.-Y., Cheng, S., Njiru, L. N., Mwavua, S. M., Obondo, A. A., & Kumar, M. (2025). Apply Machine Learning to Predict Risk for Adolescent Depression in a Cohort of Kenyan Adolescents. *Healthcare*, *13*(20), 2620. https://doi.org/10.3390/healthcare13202620

Dong, Y., Wen, H., Lu, C., Li, J., & Zheng, Q. (2025). Predicting depression risk with machine learning models: identifying familial, personal, and dietary determinants. *BMC Psychiatry*, *25*(1), 883. https://doi.org/10.1186/s12888-025-07182-8

Fassi, L., Thomas, K., Parry, D. A., Leyland-Craggs, A., Ford, T. J., & Orben, A. (2024). Social Media Use and Internalizing Symptoms in Clinical and Community Adolescent Samples: A Systematic Review and Meta-Analysis. *JAMA Pediatrics*, *178*(8), 814–822. https://doi.org/10.1001/jamapediatrics.2024.2078

Gohari, M. R., & others. (2024). Using random forest to identify correlates of depression symptoms among adolescents. *Social Psychiatry and Psychiatric Epidemiology*, *59*, 2063–2071. https://doi.org/10.1007/s00127-024-02695-1

Ke, H., Xu, A., Zhou, H., Chen, J., Wu, W., He, Q., & Cao, H. (2025). Machine learning models of depression in middle-aged and older adults with cardiovascular metabolic diseases. *Journal of Affective Disorders*, *387*, 119494. https://doi.org/10.1016/j.jad.2025.119494

Khalaf, A. M., Alubied, A. A., Khalaf, A. M., & Rifaey, A. A. (2023). The Impact of Social Media on the Mental Health of Adolescents and Young Adults: A Systematic Review. *Cureus*, *15*(8), e42990. https://doi.org/10.7759/cureus.42990

Liu, X., & others. (2025). Identification of depressive symptoms in adolescents using machine learning combining childhood and adolescence features. *BMC Public Health*, *25*, 264. https://doi.org/10.1186/s12889-025-21506-z

LoParo, D., Matos, A. P., Arnarson, E. Ö., & Craighead, W. E. (2025). Enhancing prediction of major depressive disorder onset in adolescents: A machine learning approach. *Journal of Psychiatric Research*, *182*, 235–242. https://doi.org/10.1016/j.jpsychires.2025.01.007

Mardini, M. T., & others. (2025). Identifying Adolescent Depression and Anxiety Through Real-World Data and Social Determinants of Health: Machine Learning Model Development and Validation. *JMIR Mental Health*, *12*, e66665. https://doi.org/10.2196/66665

Nagata, J. M., & others. (2025). Social Media Use and Depressive Symptoms During Early Adolescence. *JAMA Network Open*, *8*(5), e2511704. https://doi.org/10.1001/jamanetworkopen.2025.11704

Schønning, V., Hjetland, G. J., Aarø, L. E., & Skogen, J. C. (2020). Social Media Use and Mental Health and Well-Being Among Adolescents - A Scoping Review. *Frontiers in Psychology*, *11*, 1949. https://doi.org/10.3389/fpsyg.2020.01949

Twivy, E., Freeman, D., Anderson, C., Loe, B. S., & Waite, F. (2025). The social media scale for depression in adolescence. *International Journal of Adolescence and Youth*, *30*(1). https://doi.org/10.1080/02673843.2025.2450425

Valkenburg, P. M., Meier, A., & Beyens, I. (2022). Social media use and its impact on adolescent mental health: An umbrella review of the evidence. *Current Opinion in Psychology*, *44*, 58–68. https://doi.org/10.1016/j.copsyc.2021.08.017

Wu, J., Wu, X., Tarimo, C. S., & Zhao, W. (2025). Network analysis of Internet addiction and depression among Chinese adolescents. *Journal of Affective Disorders*, 119–127.

Wumaierjiang, N., Yan, G., Yuan, L., Song, J., Hou, X., Li, M., Sun, L., Zhou, J., Yin, H., & Xu, G. (2025). Predicting adolescent depressive symptoms using teacher-reported textual descriptions of abnormal behaviors: a study based on machine learning. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1732682

Zeng, S. X. (2026). Contemporary adolescent depression: Multidimensional analysis from neurobiological, psychological, and social integrative perspectives. *World Journal of Psychiatry*, *16*(3), 113125. https://doi.org/10.5498/wjp.v16.i3.113125

Zhang, H., Yu, P., Liu, X., & Wang, K. (2024). Predictive factors for the development of depression in children and adolescents: a clinical study. *Frontiers in Psychiatry*, *15*, 1460801. https://doi.org/10.3389/fpsyt.2024.1460801

Zhou, Y., Xu, H., Gong, J. P., Wan, T., & others. (2024). Identifying the risk of depression in a large sample of adolescents: An artificial neural network based on random forest. *Journal of Adolescence*, *96*(7), 1485–1497. https://doi.org/10.1002/jad.12357