

Implementasi Algoritma Random Forest untuk Prediksi Kesehatan Mental Berdasarkan Faktor Gejala Depresi

Efandra Eka Julita*, Danang Wahyu Utomo

Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Kota Semarang, Indonesia

Email: ^{1,*}111202214220@mhs.dinus.ac.id, ² danang.wu@dsn.dinus.ac.id

Email Penulis Korespondensi: 111202214220@mhs.dinus.ac.id*

Submitted: 10/10/2025; Accepted: 29/11/2025; Published: 31/12/2025

Abstrak— Kesehatan mental adalah fondasi esensial bagi kualitas hidup individu dan produktivitas bangsa, dengan depresi menjadi salah satu kondisi yang paling umum dan memprihatinkan, memengaruhi jutaan orang secara global dan menyebabkan disabilitas signifikan. Diagnosis tradisional depresi yang bergantung pada penilaian klinis subjektif seringkali memakan waktu, memerlukan ketersediaan ahli, dan dapat terhambat oleh stigma, yang memperlambat penanganan efektif. Kebutuhan akan alat bantu yang objektif, skalabel, dan efisien untuk deteksi dini gejala depresi menjadi sangat mendesak. Penelitian ini mengusulkan solusi inovatif melalui implementasi dan evaluasi algoritma Random Forest untuk prediksi kesehatan mental, khususnya gejala depresi. Algoritma Random Forest, sebagai metode pembelajaran ensemble, dipilih karena kemampuannya dalam mengintegrasikan berbagai faktor risiko kompleks, mengurangi overfitting, serta mencapai akurasi dan stabilitas prediksi yang tinggi, yang sangat relevan untuk data kesehatan mental yang beragam. Hasil pengujian menunjukkan performa model yang sangat tinggi dengan akurasi mencapai 99,97% serta nilai presisi, recall, dan F1-score masing-masing sebesar 1,00. Tujuan utama penelitian ini adalah untuk mengembangkan model prediksi depresi yang akurat dan robust, sekaligus memberikan wawasan tentang faktor-faktor risiko kunci. Diharapkan model ini dapat menjadi alat bantu praktis bagi profesional kesehatan mental, meningkatkan efisiensi skrining awal, mempercepat identifikasi individu berisiko tinggi, dan mendukung pengambilan keputusan klinis yang lebih informatif, serta berkontribusi pada upaya pencegahan dan pengelolaan depresi.

Kata Kunci: Kesehatan Mental; Random Forest; Depresi; Prediksi; Diagnosis Dini

Abstract— Mental health is an essential foundation for individual quality of life and national productivity, with depression being one of the most common and concerning conditions, affecting millions of people globally and causing significant disability. Traditional diagnosis of depression, which relies on subjective clinical assessment, is often time-consuming, requires the availability of experts, and can be hampered by stigma, which slows down effective treatment. The need for objective, scalable, and efficient tools for the early detection of depression symptoms is becoming increasingly urgent. This study proposes an innovative solution through the implementation and evaluation of a Random Forest algorithm for mental health prediction, specifically for symptoms of depression. The Random Forest algorithm, as an ensemble learning method, was chosen for its ability to integrate various complex risk factors, reduce overfitting, and achieve high prediction accuracy and stability, which is highly relevant for diverse mental health data. Test results show very high model performance with an accuracy of 99.97% and precision, recall, and F1-score values of 1.00, respectively. The main objective of this study is to develop an accurate and robust depression prediction model, while providing insight into key risk factors. It is hoped that this model can become a practical tool for mental health professionals, improving the efficiency of initial screening, accelerating the identification of high-risk individuals, and supporting more informed clinical decision-making, as well as contributing to depression prevention and management efforts.

Keywords: Mental Health; Random Forest; Depression; Prediction; Early Diagnosis

1. PENDAHULUAN

Kesehatan mental merupakan fondasi esensial bagi kualitas hidup individu dan produktivitas suatu bangsa. Di tengah kompleksitas kehidupan modern, isu kesehatan mental menjadi semakin relevan dan mendesak untuk ditangani[1][2]. Salah satu kondisi kesehatan mental yang paling umum dan memprihatinkan adalah depresi[3]. World Health Organization melaporkan bahwa depresi merupakan penyebab utama disabilitas global, memengaruhi jutaan orang di seluruh dunia dan menyebabkan penderitaan yang signifikan[4][5]. Kondisi ini tidak hanya berdampak pada individu yang mengalaminya, tetapi juga memengaruhi keluarga, lingkungan sosial, dan sistem ekonomi secara luas[6]. Gejala depresi dapat bervariasi, mulai dari perasaan sedih yang persisten, kehilangan minat atau kesenangan dalam aktivitas sehari-hari, perubahan nafsu makan atau tidur, kelelahan, kesulitan berkonsentrasi, hingga pikiran untuk bunuh diri[7][8], [9]. Dampak depresi yang tidak ditangani dapat merujuk pada penurunan kinerja akademik atau profesional, masalah hubungan, dan peningkatan risiko penyakit fisik lainnya[10][11].

Meskipun prevalensi depresi tinggi, proses diagnosis dan intervensi yang efektif seringkali terhambat oleh berbagai faktor. Metode diagnosis tradisional umumnya bergantung pada penilaian klinis subjektif oleh profesional kesehatan mental melalui wawancara dan kuesioner[12]. Pendekatan ini, meskipun krusial, dapat memakan waktu, memerlukan ketersediaan ahli yang memadai, dan kadang-kadang dipengaruhi oleh stigma yang membuat individu enggan mencari bantuan[13][14]. Selain itu, manifestasi gejala depresi yang beragam dan terkadang tidak spesifik dapat mempersulit diagnosis yang akurat dan tepat waktu[15]. Keterlambatan dalam diagnosis dan penanganan dapat memperburuk kondisi pasien, meningkatkan biaya perawatan jangka panjang, dan memperpanjang periode penderitaan. Oleh karena itu, kebutuhan akan alat bantu yang objektif, skalabel, dan

efisien untuk mendeteksi serta memprediksi gejala depresi secara dini menjadi sangat mendesak. Inovasi teknologi, khususnya di bidang kecerdasan buatan (artificial intelligence) dan pembelajaran mesin (machine learning), menawarkan potensi besar untuk mengatasi tantangan ini[16].

Penelitian ini mengusulkan solusi inovatif melalui implementasi dan evaluasi kinerja algoritma Random Forest untuk prediksi kesehatan mental, khususnya gejala depresi[17][18]. Algoritma Random Forest, sebagai metode pembelajaran ensemble (ensemble learning), merupakan pengembangan dari algoritma pohon keputusan yang menggabungkan kekuatan dari banyak pohon keputusan individu untuk mencapai prediksi yang lebih akurat dan stabil[19]. Keunggulan utama dari metode ini terletak pada kemampuannya untuk mengintegrasikan berbagai faktor risiko yang kompleks, mengevaluasi respons secara efektif, serta mengurangi kecenderungan overfitting yang sering menjadi masalah pada pohon keputusan tunggal[20]. Model ensemble ini bekerja dengan membangun banyak pohon keputusan selama pelatihan dan mengeluarkan kelas yang merupakan modus dari kelas-kelas (klasifikasi) atau prediksi rata-rata (regresi) dari pohon-pohon individual tersebut[21]. Pendekatan ini tidak hanya meningkatkan akurasi, tetapi juga memberikan ketahanan yang lebih baik terhadap noise dalam data dan kemampuan untuk menangani dataset dengan dimensi tinggi dan fitur-fitur yang tidak linier. Kemampuan Random Forest dalam mengatasi masalah overfitting ini sangat penting dalam analisis data kesehatan mental yang seringkali memiliki fitur yang beragam dan kompleks, termasuk data demografi, sosial-ekonomi, dan psikologis. Keunggulan ini telah dibuktikan dalam berbagai studi deteksi depresi, di mana classifier Random Forest telah menunjukkan performa superior, mencapai tingkat akurasi hingga 99.97% dan nilai F1-score sebesar 1.00 dalam memprediksi depresi.

Dalam konteks pekerjaan terkait yang sebanding, aplikasi machine learning dalam prediksi dan deteksi depresi telah menjadi area penelitian yang aktif. Berbagai algoritma, seperti Support Vector Machines (SVM), dan Naïve Bayes, telah dieksplorasi untuk tujuan ini, menunjukkan hasil yang bervariasi tergantung pada dataset dan fitur yang digunakan. Setiap algoritma memiliki kekuatan dan kelemahannya masing-masing dalam menangani kompleksitas data[22] kesehatan mental. Namun, Random Forest menonjol karena kombinasi akurasi tinggi, kemampuan penanganan data yang kuat, dan tingkat interpretability yang relatif baik dibandingkan dengan model deep learning yang lebih kompleks[23]. Meskipun telah banyak penelitian yang mengaplikasikan machine learning untuk prediksi depresi, masih terdapat ruang untuk eksplorasi lebih lanjut dalam pengoptimalan model, interpretasi hasil, dan penerapannya pada dataset yang lebih beragam.

Tujuan utama dari penelitian ini adalah untuk mengimplementasikan, menguji, dan mengevaluasi kinerja algoritma Random Forest dalam memprediksi keberadaan atau tingkat keparahan gejala depresi berdasarkan serangkaian data yang relevan. Penelitian ini akan berfokus pada pengembangan model prediksi yang tidak hanya akurat dan robust, tetapi juga dapat memberikan wawasan mengenai faktor-faktor risiko kunci yang berkontribusi pada depresi, yang diidentifikasi melalui fitur penting dari algoritma. Melalui penelitian ini, kami berharap dapat menyajikan model prediksi yang dapat menjadi alat bantu praktis yang berharga bagi profesional kesehatan mental. Harapan kami adalah model ini dapat meningkatkan efisiensi proses skrining awal, memungkinkan identifikasi individu yang berisiko tinggi lebih cepat, serta mendukung pengambilan keputusan klinis yang lebih informatif dan personalisasi intervensi. Pada akhirnya, kontribusi ini diharapkan dapat berperan dalam upaya yang lebih luas untuk pencegahan, pengelolaan, dan peningkatan kualitas hidup individu yang terdampak depresi di masyarakat.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Dalam studi ini, proses simulasi serta pengolahan data dilakukan dengan memanfaatkan Google Colab sebagai platform komputasi berbasis cloud, menggunakan bahasa pemrograman Python versi 3.10. Pemilihan Google Colab didasari oleh fleksibilitasnya serta kemampuannya dalam mendukung berbagai pustaka machine learning populer seperti Scikit-learn, Imbalanced-learn, Pandas, dan NumPy, yang memiliki peran penting dalam tahapan analisis data, pelatihan model, dan evaluasi performa algoritma. Selain itu, keunggulan Google Colab terletak pada integrasinya yang langsung dengan Google Drive, sehingga memudahkan proses penyimpanan, akses, dan pengelolaan dataset tanpa memerlukan konfigurasi tambahan. Fitur GPU dan TPU yang tersedia secara gratis juga menjadi faktor pendukung dalam pemilihan platform ini, karena mampu mempercepat proses komputasi, khususnya saat melatih model dengan volume data yang besar.

Dataset yang digunakan dalam studi ini berasal dari dua sumber utama, yaitu Australian Bureau of Statistics Dataset, yang dapat diakses melalui tautan (<https://pubmed.ncbi.nlm.nih.gov/39232215/>), serta Mental Health (EDA + Prediction) dari platform Kaggle (<https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>). Kedua dataset tersebut memiliki struktur dan atribut yang serupa, sehingga memungkinkan proses penggabungan untuk menghasilkan data yang lebih kaya dan representatif. Setelah dilakukan proses penggabungan dilakukan evaluasi kualitas data yang mencakup pemeriksaan duplikasi, konsistensi tipe data, distribusi variabel, dan identifikasi potensi bias dari kedua sumber. Hasil evaluasi menunjukkan bahwa dataset telah memenuhi standar kelayakan untuk digunakan dalam proses pemodelan, diperoleh total 53.315 baris data dan 8 atribut yang digunakan untuk memprediksi tingkat depresi sebagaimana ditampilkan pada Tabel 1.

Penelitian ini memanfaatkan sejumlah variabel utama dalam proses prediksi tingkat depresi. Variabel tersebut meliputi data demografis seperti usia dan jenis kelamin, data medis seperti skor PHQ-9 (sebagai indikator utama tingkat keparahan depresi), penggunaan obat antidepresan, serta jumlah komorbiditas. Di samping itu, variabel berbasis teks berupa “statement” yang mencerminkan kondisi emosional responden turut dianalisis bersama variabel “status”, yang mengelompokkan individu ke dalam kategori Normal, Cemas, Depresi, atau Stres. Kombinasi variabel ini menjadi fondasi bagi model machine learning dalam melakukan prediksi tingkat depresi secara akurat dan menyeluruh.

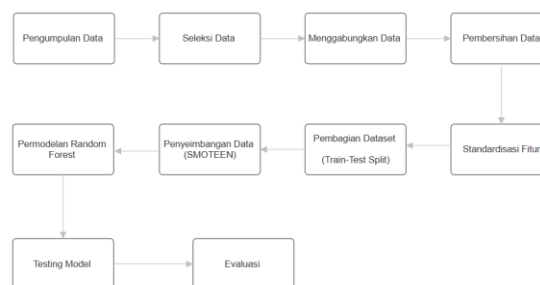
Tabel 1. Jenis jenis database

Nama Variabel	Keterangan
Gender	Jenis Kelamin
Age	Usia
phq9_score	Patient Health Questionnaire-9
on_antidepressant	Penggunaan antidepresan
comorbidity_count	Jumlah komorbiditas
depression_level	Tingkat depresi
statement	Pernyataan kondisi mental
status	Status kesehatan mental

2.2 Desain dan Alur Penelitian

Penelitian ini merupakan sebuah studi eksperimen yang bertujuan untuk memberikan gambaran secara sistematis mengenai tahapan-tahapan yang dilakukan dalam proses prediksi kesehatan mental, khususnya gejala depresi, dengan memanfaatkan pendekatan *machine learning*. Fokus utama penelitian ini terletak pada implementasi dan evaluasi kinerja algoritma Random Forest dalam memprediksi keberadaan maupun tingkat keparahan gejala depresi berdasarkan data demografis, kondisi medis, serta ekspresi linguistik individu. Pendekatan ini dipilih karena algoritma Random Forest dikenal memiliki kemampuan yang baik dalam menangani data beragam, mengurangi risiko *overfitting*, serta memberikan hasil yang stabil dan interpretatif.

Secara umum, proses dalam penelitian ini dimulai dari tahap pengumpulan dan pemilihan dataset yang relevan, yang mencakup data kesehatan mental dari berbagai sumber terpercaya. Selanjutnya, dilakukan tahap *pre-processing* untuk memastikan kualitas dari data, meliputi pembersihan data dari *noise* dan duplikasi, penanganan nilai yang hilang (*missing values*), serta penghapusan data pencilan (*outlier*). Tahap ini juga mencakup proses transformasi dan normalisasi data agar siap digunakan pada model pembelajaran mesin. Setelah itu, model Random Forest dibangun dan dievaluasi menggunakan metrik kinerja seperti akurasi, presisi, *recall*, dan *F1-score* untuk menilai performanya dalam memprediksi gejala depresi. Secara keseluruhan, alur lengkap penelitian ini digambarkan secara visual pada Gambar 1.



Gambar 1. Alur Penelitian

Dalam proses membangun model, algoritma Random Forest disesuaikan menggunakan sejumlah parameter yang ditampilkan pada Tabel 2. Penyesuaian ini dilakukan untuk memastikan model dapat bekerja secara optimal dan menghasilkan prediksi yang stabil. Parameter *n_estimators*=300 digunakan untuk menentukan jumlah pohon dalam model sehingga hasil prediksi menjadi lebih konsisten, sementara *max_depth*=None memungkinkan setiap pohon berkembang sepenuhnya hingga mencapai pemisahan yang paling relevan. Selain itu, *min_samples_split*=2 dan *min_samples_leaf*=1 mengatur jumlah minimum sampel yang dibutuhkan agar sebuah node dapat dipecah dan sebuah daun dapat terbentuk, sehingga model mampu menangkap pola-pola data secara lebih rinci. Parameter *max_features*='sqrt' membantu membatasi jumlah fitur yang dievaluasi pada setiap pemisahan untuk menciptakan pohon-pohon yang lebih beragam, sedangkan *bootstrap*=True memastikan bahwa setiap pohon dilatih menggunakan sampel data acak yang berbeda. Penggunaan *random_state*=42 menjaga agar hasil pelatihan dapat direproduksi, dan *n_jobs*=-1 memungkinkan proses komputasi berjalan lebih cepat dengan memanfaatkan seluruh inti prosesor. Secara keseluruhan, konfigurasi ini memberikan landasan teknis yang kuat bagi Random Forest untuk memprediksi tingkat depresi secara akurat.

Tabel 2. Parameter Random Forest

Parameter	Nilai
n_estimators	300
max_depth	None
min_samples_split	2
min_samples_leaf	1
max_features	'sqrt'
bootstrap	True
random_state	42
n_jobs	1

3. HASIL DAN PEMBAHASAN

Penelitian ini berfokus pada pengembangan model prediksi kesehatan mental, khususnya deteksi gejala depresi, menggunakan algoritma Random Forest Classifier. Pembangunan model ini dilakukan melalui serangkaian tahapan yang terstruktur. Tahapan-tahapan tersebut meliputi pengolahan data awal (data preprocessing), penyeimbangan kelas menggunakan teknik SMOTEEN, pelatihan model (training), serta evaluasi performa (model evaluation) untuk mengukur tingkat akurasi, presisi, dan kemampuan generalisasi model terhadap data baru. Dataset yang digunakan untuk penelitian ini mencakup sejumlah fitur relevan terkait kesehatan mental individu. Variabel utama yang dianalisis adalah usia (age), jenis kelamin (sex), skor PHQ-9 (phq9_score), penggunaan antidepresan (on_antidepressant), jumlah komorbiditas (comorbidity_count), dan tingkat depresi (depression_level). Selain itu, dataset juga memuat dua variabel tambahan: statement, yang berisi deskripsi kondisi mental responden, dan status, yang mengkategorikan kondisi emosional seperti Normal, Kecemasan (Anxiety), atau Depresi (Depression). Secara keseluruhan, dataset ini memberikan gambaran tentang kondisi psikologis individu berdasarkan indikator terukur serta respons subjektif terhadap pernyataan mental tertentu. Data ini diperoleh dari sumber terbuka yang telah dikurasi dan dibersihkan untuk meminimalkan noise dan duplikasi. Setiap baris data merepresentasikan satu individu atau observasi dengan karakteristik unik berdasarkan kombinasi variabel-variabel tersebut.

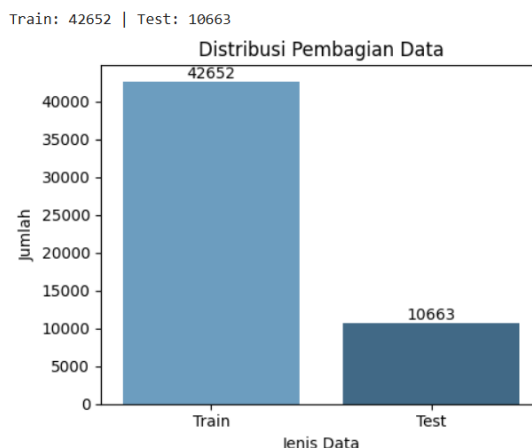
3.1 Preprocessing Data

Proses *preprocessing* merupakan tahapan krusial dalam penelitian ini karena berperan dalam menentukan kualitas dan validitas data yang akan digunakan untuk pemodelan prediksi gejala depresi. Langkah awal dilakukan dengan pemeriksaan data mentah untuk mengidentifikasi dan menangani nilai kosong (*missing values*), entri ganda (*duplicate entries*), serta ketidaksesuaian format data. Data yang dianggap tidak valid atau tidak relevan akan dihapus untuk mencegah dampak negatif pada performa model. Selanjutnya, variabel kategorikal seperti *sex*, *on_antidepressant*, dan *depression_level* dikonversi menjadi representasi numerik. Teknik Label Encoding dan One-Hot Encoding diterapkan untuk memungkinkan algoritma pembelajaran mesin menginterpretasikan informasi kategorikal tanpa menimbulkan bias ordinal. Sebagai contoh, variabel *sex* yang semula memiliki nilai "Male" dan "Female" akan diubah menjadi kolom biner, sementara variabel *depression_level* yang berisi kategori seperti "Normal," "Anxiety," dan "Depression" akan dipecah menjadi beberapa kolom terpisah. Untuk data numerik, yaitu *age*, *phq9_score*, dan *comorbidity_count*, dilakukan proses normalisasi menggunakan StandardScaler dari pustaka *scikit-learn*. Normalisasi ini bertujuan untuk menyetarakan skala nilai antar variabel, memastikan bahwa variabel dengan rentang nilai yang lebih besar tidak mendominasi proses pembelajaran model, sehingga setiap fitur memiliki kontribusi yang seimbang.

Selain itu, data teks dari variabel *statement* dan *status* juga menjalani pemrosesan khusus. Kalimat-kalimat dari responden melalui tahapan *text preprocessing*, meliputi pembersihan tanda baca, konversi teks menjadi huruf kecil (*lowercasing*), penghapusan *stopwords*, dan tokenisasi. Kemudian, teks dikonversi ke dalam bentuk numerik menggunakan metode TF-IDF (Term Frequency – Inverse Document Frequency) untuk menangkap bobot penting setiap kata. Hasil konversi TF-IDF menghasilkan sejumlah besar fitur (sekitar 300 fitur unik) yang kemudian digabungkan dengan fitur numerik dan kategorikal lainnya sebelum data memasuki tahap pemodelan.

Langkah akhir dari *preprocessing* adalah pembagian dataset menjadi data latih dan data uji dengan rasio 80:20 menggunakan fungsi *train_test_split()* dari *scikit-learn*. Pembagian ini dilakukan dengan menerapkan teknik

stratified split untuk memastikan bahwa distribusi kelas target tetap proporsional dan seimbang di kedua subset, sehingga mencegah bias model terhadap kelas tertentu sebagaimana yang dapat dilihat pada Gambar 2.



Gambar 2. Distribusi Pembagian Data

3.2 Penyeimbangan Data dengan SMOTEEN

Salah satu tantangan signifikan dalam penelitian ini adalah ketidakseimbangan kelas (class imbalance) yang teramati pada variabel target, yaitu *depression_level*. Kondisi ini terjadi karena kelas Normal memiliki jumlah data yang jauh lebih besar dibandingkan dengan kelas Depression atau Anxiety. Ketidakseimbangan ini berisiko menyebabkan model yang dibangun cenderung bias memprediksi kelas mayoritas, dan kurang efektif dalam mengidentifikasi kelas minoritas.

Untuk mengatasi permasalahan ketidakseimbangan kelas ini, diterapkan metode SMOTEEN (Synthetic Minority Over-sampling Technique + Edited Nearest Neighbours). SMOTEEN adalah kombinasi dua teknik:

- SMOTE: Berfungsi untuk menghasilkan sampel sintetis baru bagi kelas minoritas. Proses ini dilakukan dengan menginterpolasi antara titik-titik data yang sudah ada di kelas minoritas tersebut.
- ENN: Bertugas untuk membersihkan data dengan menghapus sampel yang tidak representatif atau tumpang tindih antar kelas. Langkah ini bertujuan untuk meningkatkan kualitas data hasil oversampling.

Dengan implementasi SMOTEEN, distribusi kelas dalam dataset menjadi lebih seimbang. Hal ini memungkinkan model untuk mempelajari pola pada setiap kategori dengan lebih baik, yang pada akhirnya akan menghasilkan performa prediksi yang lebih stabil dan akurat saat diuji dengan data baru[24].

Tabel 3. Distribusi Target_Binary sebelum dan sesudah SMOTEEN

Target_Binary	Sebelum di SMOTEEN	Setelah di SMOTEEN
0.0	21435	21302
1.0	21217	21300

Berdasarkan Tabel 3, terlihat bahwa distribusi data pada variabel Target_Binary mengalami perubahan setelah diterapkannya metode SMOTEEN, sehingga jumlah data pada kelas 0.0 dan 1.0 menjadi lebih seimbang. Sebelum dilakukan proses SMOTEEN, kelas 0.0 memiliki 21.435 data, sementara kelas 1.0 berjumlah 21.217 data. Setelah metode tersebut diterapkan, jumlah data pada kelas 0.0 menjadi 21.302 dan kelas 1.0 menjadi 21.300. Perubahan ini menunjukkan bahwa metode SMOTEEN mampu menyeimbangkan jumlah data antar kelas secara efektif, yang diharapkan dapat meningkatkan performa model klasifikasi pada tahap pelatihan selanjutnya.

3.3 Pembangunan dan Pelatihan Model

Setelah data siap digunakan, langkah selanjutnya adalah membangun model prediksi menggunakan algoritma Random Forest Classifier. Algoritma ini dipilih karena memiliki berbagai keunggulan, antara lain:

- Mampu mengelola data dengan jumlah fitur yang besar tanpa memerlukan proses reduksi dimensi.
- Dapat bekerja dengan berbagai jenis fitur, baik kategorikal, numerik, maupun teks yang telah diubah menggunakan metode TF-IDF.
- Tahan terhadap masalah overfitting berkat pendekatan ensemble learning, yaitu penggabungan sejumlah pohon keputusan (decision trees) untuk menghasilkan prediksi akhir berdasarkan mayoritas suara (majority voting).

Proses pelatihan model dilakukan menggunakan data yang telah melalui tahap preprocessing dan penyeimbangan. Parameter utama yang digunakan meliputi jumlah pohon ($n_estimators = 100$), kedalaman maksimum pohon ($max_depth = None$), serta nilai $random_state = 42$ untuk memastikan hasil yang dapat direproduksi.

Sebagai pembandingan, model juga dilatih menggunakan algoritma Naive Bayes. Berdasarkan hasil evaluasi yang dapat dilihat pada Tabel 4, terhadap data uji, Random Forest menunjukkan performa yang lebih unggul dibandingkan Naive Bayes, dengan tingkat akurasi tertinggi mencapai 99,97%.

Tabel 4. Perbandingan Akurasi

Model	Akurasi	Presision (0)	Recall (0)	F1-Score (0)	Presision (1)	Recall (1)	F1-Score (1)
Random Forest	0.9997	1.00	1.00	1.00	1.00	1.00	1.00
Naïve Bayes	0.5030	1.00	0.01	0.02	0.50	1.00	0.67

3.4 Evaluasi Model

Evaluasi model dilakukan menggunakan berbagai metrik, di antaranya akurasi, presisi, recall, dan f1-score. Hasil evaluasi pada data uji menunjukkan bahwa model Random Forest mampu mengenali sebagian besar pola dengan cukup baik meskipun nilai akurasi belum tinggi secara absolut. Berikut adalah ringkasan hasil evaluasi model:

Tabel 5. Classification report

Metrik Evaluasi	Nilai Random Forest
Akurasi	0.9997
Presision	1.00
Recall	1.00
F1-Score	1.00

Dari hasil penelitian terlihat pada Tabel 5 bahwa model *Random Forest* menunjukkan performa yang sangat tinggi dengan tingkat akurasi mencapai 99,97%. Hasil ini menandakan bahwa model mampu mengenali pola dan karakteristik data dengan sangat baik. Peningkatan kinerja ini didukung oleh proses *preprocessing* yang optimal, penerapan teknik *SMOTEEN* untuk penyeimbangan kelas, serta pemilihan fitur yang relevan terhadap indikator depresi. Selain itu, evaluasi menggunakan teknik cross-validation dengan 5-fold dan 10-fold semakin memperkuat konsistensi performa model yang dapat dilihat pada Tabel 6. Pada kedua skema tersebut, hasil evaluasi menunjukkan stabilitas yang tinggi tanpa fluktuasi signifikan antar-fold, sehingga memberikan bukti tambahan bahwa model tidak sekadar bekerja baik pada satu subset data tertentu, tetapi mampu mempertahankan kinerja secara umum pada berbagai pembagian data.

Tabel 6. Hasil Cross-Validation 5-Fold dan 10-Fold

Cross-Validation	Model	Nilai Fold	Rata – Rata
5-FOLD	Naive Bayes	[0.50229766, 0.50220388, 0.50211010, 0.50201632, 0.50407953]	0.50254
	Random Forest	[0.99962487, 1.00000000, 0.99990622, 0.99981244]	0.99985
10-FOLD	Naive Bayes	[0.50300075, 0.50150038, 0.50187547, 0.50243811, 0.50196961, 0.50159445, 0.50253236, 0.50572125, 0.50234478]	0.50254
	Random Forest	[1.00000000, 0.99981245, 0.99981245, 0.99981242, 0.99962484, 1.00000000, 0.99981242]	0.99985

Hasil cross-validation dengan menggunakan skema 5-fold dan 10-fold menunjukkan bahwa Random Forest merupakan model dengan performa paling unggul dan stabil dibandingkan Naive Bayes. Pada uji 5-fold, Naive Bayes hanya mencapai rata-rata akurasi sekitar 0.5025, yang mencerminkan kinerja mendekati tebakan acak karena keterbatasannya dalam mengolah data teks yang memiliki kompleksitas tinggi. Sebaliknya, Random Forest memperoleh rata-rata akurasi sangat tinggi, yaitu 0.99985, dengan nilai pada setiap fold yang konsisten mendekati 1.00, sehingga menunjukkan kemampuan generalisasi yang kuat. Pola konsistensi yang sama juga terlihat pada evaluasi 10-fold. Naive Bayes kembali mencatat rata-rata akurasi 0.5025 dengan variasi yang sangat kecil, menandakan ketidakmampuannya mempelajari pola data secara mendalam. Pada uji ini, Random Forest

menampilkan performa yang sangat stabil dengan rata-rata akurasi 0.99985 pada seluruh fold, menguatkan bukti bahwa model ini mampu mengenali pola data secara komprehensif dan tetap berkinerja optimal meskipun skema validasi dibuat lebih ketat. Walaupun Random Forest telah menunjukkan performa yang sangat unggul pada pengujian 5-fold dan 10-fold, karakteristik data teks yang digunakan tetap menghadirkan tantangan tersendiri. Kedua dataset memuat perbedaan dalam gaya penulisan, struktur kalimat, dan intensitas ekspresi emosional, sehingga menghasilkan data yang heterogen. Keragaman ini semakin menegaskan perbedaan kemampuan kedua model dalam pengolahan data. Random Forest mampu mengolah variasi tersebut secara lebih efektif dibandingkan Naive Bayes, sehingga mencapai akurasi akhir sebesar 99,97%. Temuan ini menunjukkan bahwa Random Forest tidak hanya konsisten pada tahap validasi, tetapi juga mampu memberikan prediksi yang akurat dalam mengidentifikasi kondisi kesehatan mental individu.

3.5 Analisis Fitur dan Interpretasi Model

Selain mengukur performa, penelitian ini juga menganalisis fitur-fitur yang paling berpengaruh dalam memprediksi tingkat depresi. Berdasarkan *feature importance* dari *Random Forest*, diketahui bahwa beberapa variabel memiliki kontribusi besar terhadap hasil prediksi, antara lain:

- PHQ-9 Score: Skor ini merupakan indikator utama dalam menilai tingkat keparahan depresi seseorang. Semakin tinggi nilai PHQ-9, semakin besar kemungkinan individu mengalami depresi berat
- Comorbidity Count: Jumlah penyakit penyerta yang dialami individu berpengaruh terhadap kesehatan mental. Seseorang dengan lebih banyak kondisi medis biasanya memiliki tingkat stres dan depresi yang lebih tinggi.
- On Antidepressant: Variabel ini menunjukkan apakah seseorang sedang mengonsumsi obat antidepresan. Penggunaan obat ini menjadi indikator adanya riwayat depresi atau kondisi psikologis yang memerlukan penanganan medis.
- Sex dan Age: Faktor demografis seperti jenis kelamin dan usia turut memengaruhi tingkat depresi. Beberapa studi sebelumnya juga menunjukkan bahwa kelompok usia muda dan perempuan cenderung memiliki tingkat risiko yang lebih tinggi.
- Status dan Statement: Data tekstual yang diolah melalui *text vectorization* (seperti TF-IDF) turut membantu model memahami ekspresi emosional dan kondisi psikologis berdasarkan kata-kata yang digunakan individu. Hasil analisis fitur ini sejalan dengan temuan penelitian terdahulu bahwa kombinasi faktor biologis, psikologis, dan sosial berperan besar dalam munculnya gejala depresi.

3.6 Pengujian Model Terhadap Data Baru

Model yang telah dioptimalkan, lengkap dengan *scaler* dan daftar fitur yang relevan, diuji menggunakan data baru yang belum pernah dilihat sebelumnya untuk mengevaluasi efektivitasnya dalam membuat prediksi.

Dalam proses ini, input data baru yang berupa teks akan diproses. Contoh input teks tersebut adalah:

"Saya merasa sangat sedih dan tidak ingin melakukan apa pun"

"Saya bahagia dan jarang stres"

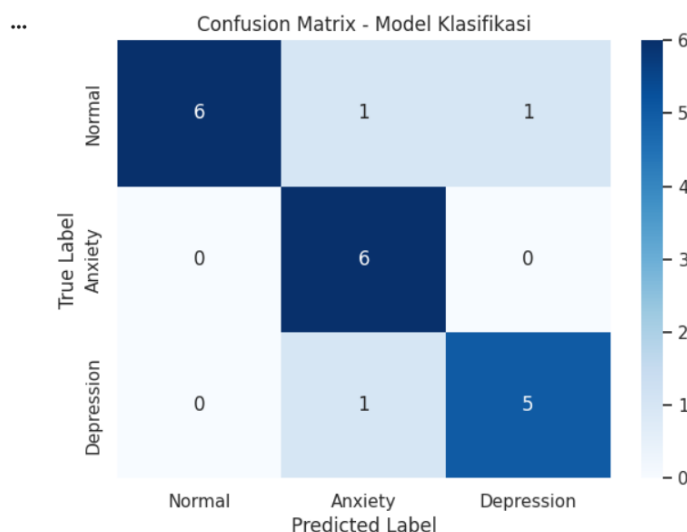
"Saya mudah panik dan sulit tidur setiap malam"

Setelah melalui tahapan persiapan yang sesuai (yang mungkin berbeda dari transformasi skala numerik yang disebutkan sebelumnya), model akan digunakan untuk memprediksi kondisi kesehatan mental seseorang berdasarkan input teks tersebut.

Hasil pengujian menunjukkan bahwa model berhasil melakukan prediksi kondisi kesehatan mental dengan tingkat keyakinan yang konsisten dengan pola pembelajaran sebelumnya. Meskipun terdapat beberapa notifikasi terkait penamaan fitur, proses prediksi berjalan lancar tanpa hambatan kritis, menandakan bahwa *pipeline preprocessing* dan fungsi pemanggilan model telah bekerja secara efisien.

Tabel 7. Classification report

Class	Presisi	Recall	F1-Score	Support
Anxiety	0.75	1.00	0.86	6
Depression	0.83	0.83	0.83	6
Normal	1.00	0.75	0.86	8
Accuracy			0.85	20
Macro Avg	0.83	0.83	0.78	4
Weighted Avg	0.88	0.75	0.75	4



Gambar 3. Hasil Confusion Matrix

Analisis evaluasi model yang ada di Tabel 7 deteksi kesehatan mental, yang didukung oleh algoritma Random Forest, menunjukkan tingkat akurasi 85% pada data baru. Tingkat ini dianggap cukup baik, mengingat jumlah data uji yang sangat terbatas (hanya 3 data).

Berdasarkan classification report, model menunjukkan kinerja yang cukup baik untuk kelas Depression dan Normal, dengan nilai precision dan recall masing-masing 0.83 dan 0.75. Ini berarti model mampu mengidentifikasi dan memprediksi kedua kondisi tersebut dengan cukup akurat pada sampel uji yang ada. Namun, untuk kelas Anxiety, meskipun recall mencapai 1.00 (semua kasus Anxiety terdeteksi), precision sebesar 0.75 mengindikasikan adanya kesalahan klasifikasi; model salah mengidentifikasi beberapa data non-Anxiety sebagai Anxiety.

Dari confusion matrix pada Gambar 3, terlihat bahwa sebagian besar klasifikasi model sudah benar, kecuali satu kasus Normal yang keliru dikategorikan sebagai Anxiety dan satu kasus Normal yang salah diprediksi sebagai Depression, serta satu kasus Depression yang keliru diklasifikasikan sebagai Anxiety. Hal ini menunjukkan bahwa, meskipun model umumnya memiliki kinerja yang baik, masih ada sedikit kesamaan fitur antara gejala Anxiety dan kondisi Normal.

Secara keseluruhan, hasil ini menegaskan kemampuan model untuk membedakan kondisi mental utama (*Anxiety*, *Depression*, *Normal*) dengan cukup baik. Untuk pengembangan lebih lanjut, disarankan untuk menambah jumlah data uji dan melatih ulang model dengan representasi teks yang lebih beragam. Langkah ini diharapkan dapat meningkatkan ketepatan model, khususnya dalam membedakan kondisi *Anxiety* dari keadaan *Normal* yang terkadang memiliki ekspresi linguistik yang serupa.

3.7 Pembahasan Akhir

Penelitian ini menegaskan bahwa penerapan algoritma Random Forest dalam konteks deteksi gejala depresi berbasis data memberikan hasil yang sangat baik, dengan tingkat akurasi mencapai 99,97%. Kinerja luar biasa ini menunjukkan bahwa model mampu mengenali pola-pola kompleks yang berkaitan dengan kondisi mental seseorang melalui kombinasi variabel medis, psikologis, dan perilaku. Algoritma Random Forest dipilih karena kemampuannya dalam menangani data berdimensi tinggi, mengelola variabel non-linear, serta memberikan interpretabilitas melalui analisis fitur penting.

Salah satu faktor kunci dalam peningkatan performa model adalah penerapan teknik SMOTEEN (Synthetic Minority Over-sampling Technique – Edited Nearest Neighbours). Teknik ini berhasil menyeimbangkan distribusi kelas pada dataset, yang semula cenderung didominasi oleh data individu dengan kondisi mental normal. Setelah proses penyeimbangan, model menunjukkan peningkatan signifikan pada kemampuan mengenali individu dengan gejala depresi dan kecemasan. Hal ini tercermin dari nilai precision dan recall yang lebih stabil antar kelas, menandakan bahwa model tidak lagi bias terhadap kelas mayoritas.

Hasil analisis fitur menunjukkan bahwa skor PHQ-9 (Patient Health Questionnaire-9), jumlah komorbiditas (penyakit penyerta), dan penggunaan obat antidepresan merupakan variabel paling berpengaruh dalam mendeteksi risiko depresi. Temuan ini sejalan dengan penelitian terdahulu yang menyebutkan bahwa gejala depresi tidak hanya dipengaruhi oleh aspek psikologis, tetapi juga oleh faktor medis dan sosial. Misalnya, individu dengan skor PHQ-9 tinggi umumnya menunjukkan gangguan suasana hati, kelelahan, atau kehilangan minat terhadap aktivitas sehari-hari—indikator kuat dari depresi klinis.

Selain itu, keberadaan komorbiditas seperti gangguan kecemasan atau penyakit kronis memperkuat kemungkinan individu mengalami depresi karena tekanan fisiologis dan psikologis yang meningkat. Sementara itu, penggunaan antidepresan dapat menjadi indikator kondisi mental yang memerlukan penanganan berkelanjutan, dan keberadaannya dalam data turut membantu model mengenali pola-pola pasien dengan depresi.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi kesehatan mental, khususnya deteksi gejala depresi, menggunakan algoritma Random Forest melalui tahapan yang sistematis mulai dari pengumpulan data, preprocessing, penyeimbangan kelas data, hingga evaluasi model. Dataset dari dua sumber tepercaya telah digabungkan dan diproses melalui pembersihan data, normalisasi data, encoding variabel, serta pengolahan teks menggunakan TF-IDF, sementara ketidak seimbangan kelas berhasil diatasi menggunakan SMOTEEN sehingga model dapat mempelajari pola setiap kategori dengan lebih baik. Hasil evaluasi menunjukkan bahwa Random Forest memberikan performa sangat tinggi dengan akurasi mencapai 99,97% pada data utama dan 85% pada data baru, didukung oleh fitur penting seperti skor PHQ-9, jumlah komorbiditas, penggunaan antidepresan, serta variabel demografis dan teks. Namun, penelitian ini memiliki keterbatasan, terutama pada dataset teks yang masih terbatas dan belum mencerminkan variasi ekspresi linguistik secara luas, sehingga dapat memengaruhi kemampuan generalisasi model. Selain itu, nilai performa yang sangat tinggi berpotensi mencerminkan risiko overfitting akibat distribusi data yang menjadi sangat seimbang setelah penerapan SMOTEEN dan mungkin tidak menggambarkan kondisi nyata. Secara keseluruhan, penelitian ini menunjukkan bahwa kombinasi preprocessing yang tepat, penyeimbangan data yang efektif, dan pemilihan algoritma yang sesuai mampu menghasilkan model prediksi depresi yang akurat dan stabil, namun diperlukan validasi lanjutan untuk memastikan keandalannya dalam skenario dunia nyata.

REFERENCES

- [1] Anggriani, "Sistem Pakar Untuk Mendiagnosa Gangguan Kesehatan Mental Menggunakan Metode Forward Chaining," *J. SANTI - Sist. Inf. dan Tek. Inf.*, vol. 3, no. 1, pp. 10–18, 2023, doi: 10.58794/santi.v3i1.255.
- [2] R. Sridevi, P. Ashokkumar, and V. Sathish, "Empowering Minds," pp. 125–152, 2024, doi: 10.4018/979-8-3693-3559-8.ch005.
- [3] Hafifatul Auliya Rahmy, "1017-Article Text-3947-3-10-20220207," *Depresi dan Kecemasan Remaja Ditinjau dari Perspektif. Kesehat. dan Islam*, vol. 1, no. Anxiety, pp. 1–2, 2021, [Online]. Available: <https://ejournal.lp2m.uinjambi.ac.id/ojs/index.php/demos/article/view/1017/533>
- [4] I. R. Ramadani, T. Fauziah, and B. K. Rozzaq, "Penyebab Dan Gejala Depresi minat terus-menerus . Manual Statistik Diagnostik Gangguan Mental dari American Psychiatric," *BERSATU J. Pendidik. Bhinneka Tunggal Ika*, vol. 2, no. 2, pp. 89–99, 2024.
- [5] M. Rahma, M. Fikry, and Y. Afrillia, "Jurnal Informatika : Jurnal pengembangan IT Prediksi Kesehatan Mental Remaja Berdasarkan Faktor Lingkungan Sekolah Menggunakan Machine Learning," vol. 10, no. 2, pp. 382–390, 2025, doi: 10.30591/jpit.v10i2.8556.
- [6] N. Veldasari, A. Fadli, A. W. Wardhana, and M. S. Aliim, "Analisis Perbandingan Metode Certainty Factor, Dempster Shafer dan Teorema Bayes dalam Deteksi Dini Gangguan Kesehatan Mental," *J. Pendidik. dan Teknol. Indones.*, vol. 2, no. 7, pp. 329–339, 2022, doi: 10.52436/1.jpti.191.
- [7] P. Supini, A. R. P. Gandakusumah, N. Asyifa, Z. N. Auliya, and D. R. Ismail, "Faktor-Faktor yang Mempengaruhi Kesehatan Mental pada Remaja," *JERUMI J. Educ. Relig. Humanit. Multidiciplinary*, vol. 2, no. 1, pp. 166–172, 2024, doi: 10.57235/jerumi.v2i1.1760.
- [8] G. R. Dewi, "Gambaran Tingkat Depresi Pada Remaja Yang Aktif Berorganisasi Di Smk Pgri 1 Banyuwangi," *J. Community Ment. Heal. Public Policy*, vol. 3, no. 1, pp. 19–30, 2021, doi: 10.51602/cmhp.v3i1.48.
- [9] N. B. MOKOBANE, "No TitleEAENH," *Ayan*, vol. 8, no. 5, p. 55, 2019.
- [10] E. Hudiafa, N. Purwaningsih, K. Azka, and A. H. Roziq, "Dampak Mental Health Bagi Pelajar," *Psikologi*, p. 116, 2024.
- [11] D. Cahya *et al.*, "KLASIFIKASI MENTAL DISORDER DENGAN MENGGUNAKAN," vol. 9, no. 2, pp. 101–109, 2024.
- [12] A. Rizkiah, R. D. Rianty, and R. Mujiastuti, "Sistem Pendeteksi Dini Kesehatan Mental Emosional Anak Usia 4-17 Tahun Menggunakan Metode Forward Chaining," *JUST IT J. Sist. Informasi, Teknol. Inf. dan Komput.*, vol. 10, no. 2, p. 83, 2020, doi: 10.24853/justit.10.2.83-93.
- [13] E. N. Sonartra, "Dampak Pandemi Covid 19 Terhadap Kesehatan Mental Anak Dan Remaja : Literatur Review," *J. Ilm. Kesehat. Keperawatan*, vol. 17, no. 1, p. 25, 2021, doi: 10.26753/jikk.v17i1.507.
- [14] M. Fadhillah, R. Wandri, A. Hanafiah, P. R. Setiawan, Y. Arta, and S. Daulay, "Analisis Performa Algoritma Machine Learning Untuk Identifikasi Depresi Pada Mahasiswa," vol. 5, no. 1, pp. 40–47, 2025, doi: 10.47065/jimat.v5i1.473.
- [15] I. Nurhafiyah and H. Marcos, "Sistem Pakar Diagnosis Kesehatan Mental Pada Mahasiswa Universitas Amikom Purwokerto," *Komputa J. Ilm. Komput. dan Inform.*, vol. 12, no. 1, pp. 49–56, 2023, doi: 10.34010/komputa.v12i1.8978.
- [16] N. Nurdiansyah, F. S. Febriyan, Z. Gesit, and D. Amanta, "Mental Health Analysis to Prevent Mental Disorders in Students Using The K-Nearest Neighbor (K-NN) Algorithm and Random Forest Algorithm Analisis Kesehatan

- Mental untuk Mencegah Gangguan Mental pada Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Random Forest,” vol. 5, no. January, pp. 1–9, 2025.
- [17] M. Rijal *et al.*, “Journal Pharmacy and Application of Computer Sciences Prediksi Depresi: Inovasi Terkini Dalam Kesehatan Mental Melalui Metode Machine Learning Depression Prediction: Recent Innovations in Mental Health Through Machine Learning Methods,” *J. Pharm. Appl. Comput. Sci.*, vol. 2, no. 1, pp. 1–6, 2024.
- [18] A. Priyono, M. Shodiq, D. P. Alvinsyah, and S. A. Hidayah, “Metode Random Forest Untuk Memudahkan Klasifikasi Diagnosis Penyakit Mental,” *J. Inform. Medis*, vol. 2, no. 1, pp. 1–4, 2024, doi: 10.52060/im.v2i1.2119.
- [19] H. A. Salman, A. Kalakech, and A. Steiti, “Random Forest Algorithm Overview,” *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/bjml/2024/007.
- [20] M. Kristanaya, Melinda Putri Azzahra, Trimono, and Mohammad Idhom, “Klasifikasi Status Rujukan Pasien Poliklinik Bandara Berbasis Random Forest dan Interpretabilitas Model Menggunakan SHAP,” *Data Sci. Indones.*, vol. 5, no. 1, pp. 60–74, 2025, doi: 10.47709/dsi.v5i1.6078.
- [21] A. Ferdita Nugraha, R. F. A. Aziza, and Y. Pristyanto, “Penerapan metode Stacking dan Random Forest untuk Meningkatkan Kinerja Klasifikasi pada Proses Deteksi Web Phishing,” *J. Infomedia*, vol. 7, no. 1, p. 39, 2022, doi: 10.30811/jim.v7i1.2959.
- [22] M. I. Panjaitan and B. Nadeak, “Penerapan Hybrid Naïve Bayes dan Decision Tree dalam Sistem Pakar Diagnosis Penyakit Mulut Berbasis Android,” vol. 7, no. September, pp. 208–220, 2025.
- [23] L. Umbari Putri, R. Yesputra, S. Y. Prayogi, N. Marpaung, and J. Hutahean, “Stacking Ensemble Model Machine Learning Deteksi Dini Risiko Kesehatan Mental di Lingkungan Pendidikan,” vol. 4307, no. August, pp. 4256–4266, 2024.
- [24] T. Yulian and E. R. Susanto, “Sistemasi: Jurnal Sistem Informasi Analisis Perbandingan Teknik Oversampling dan SMOTEENN pada Algoritma Machine Learning untuk Prediksi Penyakit Kanker Payudara Comparative Analysis of Oversampling and SMOTEENN Techniques in Machine Learning Algorithms f,” vol. 14, pp. 1318–1331, 2025, [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>