JURNAL BUFFER INFORMATIKA
Volume 12 Nomor 1, April 2026
p-ISSN : 2527-4856, e-ISSN : 2614-5413
https://journal.fkom.uniku.ac.id/index.php/buffer

Klasifikasi Tingkat Depresi Menggunakan Algoritma Random Forest Berdasarkan Faktor Gaya Hidup, Demografi, dan Psikologis

Haniyah Faizah[1)] **, Oktavio Theonady**[2)]**, Syalwa Salsabillah Siregar**[3)]**, Fathoni.**[4)]**, Ibrahim,**[5)]

[1,2,3,4,5] *Fakultas Ilmu Komputer, Universitas Sriwijaya*
Jalan Srijaya Negara, Bukit Lama, Kec. Ilir Barat I, Kota Palembang, Sumatera Selatan (30139)
Email: 09031382328167@student.unsri.ac.id[1)],09031182328012@student.unsri.ac.id[2)],
09031382328139@student.unsri.ac.id[3)],fathoni@unsri.ac.id[4)],aliibrahim@unsri.ac.id[5)]

Abstrak

Depresi pada mahasiswa merupakan isu kesehatan mental yang berdampak signifikan terhadap kualitas hidup dan performa akademik. Meskipun banyak penelitian telah dilakukan, sebagian besar masih menganalisis faktor gaya hidup, demografi, dan psikologis secara terpisah serta menggunakan evaluasi model yang terbatas pada akurasi, sehingga belum mencerminkan performa model secara komprehensif. Penelitian ini bertujuan mengembangkan model klasifikasi tingkat depresi yang terintegrasi dan *robust* menggunakan algoritma Random Forest. Dataset yang digunakan berasal dari Mendeley Data berupa data survei kesehatan mental *(self-reported)* dengan 1.998 sampel. Metode penelitian mengacu pada *Knowledge Discovery in Database* (KDD) dengan tahapan seleksi fitur menggunakan *Chi-Square, preprocessing,* transformasi data, serta penanganan ketidakseimbangan kelas menggunakan SMOTE. Evaluasi model dilakukan secara komprehensif menggunakan *accuracy, precision, recall, F1-score,* dan *ROC-AUC*, serta divalidasi menggunakan *Stratified 5-Fold Cross-Validation* untuk menguji stabilitas dan generalisasi model. Hasil penelitian menunjukkan bahwa Random Forest memberikan performa terbaik dibandingkan KNN dan SVM dengan akurasi 96,17% dan ROC-AUC 0,9977, serta nilai *precision, recall,* dan *F1-score* yang tinggi dan konsisten. Variabel *Symptoms, Nervous Level*, dan *Employment Status* teridentifikasi sebagai faktor dominan dalam klasifikasi tingkat depresi. Model yang dihasilkan berpotensi digunakan sebagai dasar pengembangan sistem deteksi dini depresi yang lebih akurat dan mendukung pengambilan keputusan dalam intervensi kesehatan mental di lingkungan perguruan tinggi.

Kata Kunci : Depresi; Random Forest; Klasifikasi; Gaya Hidup; Psikologis

Abstract

Depression among college students is a mental health issue that significantly impacts quality of life and academic performance. Although numerous studies have been conducted, most still analyze lifestyle, demographic, and psychological factors separately and employ model evaluations limited to accuracy, thus failing to comprehensively reflect model performance. This study aims to develop an integrated and robust depression severity classification model using the Random Forest algorithm. The dataset used comes from Mendeley Data in the form of mental health survey data (self-reported) with 1,998 samples. The research method refers to Knowledge Discovery in Databases (KDD) with feature selection stages using Chi-Square, preprocessing, data transformation, and class imbalance handling using SMOTE. Model evaluation was conducted comprehensively using accuracy, precision, recall, F1-score, and ROC-AUC, and validated using Stratified 5-Fold Cross-Validation to test the model's stability and generalization. The results show that Random Forest performs best compared to KNN and SVM, with an accuracy of 96.17% and an ROC-AUC of 0.9977, as well as high and consistent precision, recall, and F1-score values. The variables Symptoms, Nervous Level, and Employment Status were identified as dominant factors in depression level classification. The resulting model has the potential to serve as the basis for developing a more accurate early-detection system for depression and to support decision-making regarding mental health interventions in university settings.

Keywords: Depression; Random Forest; Classification; Lifestyle; Psychological

1. PENDAHULUAN

Depresi merupakan salah satu gangguan kesehatan mental dengan prevalensi tinggi pada kelompok mahasiswa dan dewasa muda [1], [5]. Tanpa deteksi dini yang memadai, kondisi ini dapat menurunkan performa akademik, mengganggu interaksi sosial, serta berdampak signifikan terhadap kualitas hidup individu [2], [4]. Seiring berkembangnya teknologi, pendekatan berbasis *machine learning* banyak dimanfaatkan untuk mendukung proses identifikasi dan klasifikasi tingkat depresi secara objektif berbasis data. Sejumlah penelitian menunjukkan bahwa algoritma seperti Random Forest mampu menghasilkan performa klasifikasi yang tinggi dengan nilai akurasi dan *Area Under the Curve* (AUC) yang optimal [1], [3], [6]. Namun demikian, performa yang tinggi tidak serta-merta mencerminkan kemampuan generalisasi model terhadap data baru.

Dalam implementasinya, penerapan *machine learning* pada domain kesehatan mental menghadapi sejumlah tantangan metodologis yang signifikan, khususnya *overfitting*, *data leakage*, dan *bias dataset*. *Data leakage* terjadi ketika informasi dari data uji secara tidak langsung digunakan dalam proses pelatihan, sehingga menghasilkan estimasi performa yang bias dan tidak representatif terhadap kondisi nyata [13]. Selain itu, *overfitting* menyebabkan model terlalu menyesuaikan pola pada data latih sehingga menurunkan kemampuan generalisasi pada data baru. Karakteristik data kesehatan mental yang umumnya berbasis survei (*self-reported*) juga meningkatkan risiko bias, baik dalam bentuk ketidakseimbangan kelas maupun subjektivitas responden, yang dapat menyebabkan model cenderung mengoptimalkan kelas mayoritas dan mengabaikan pola penting pada kelas minoritas [14]. Kondisi ini menegaskan bahwa evaluasi model tidak dapat hanya berfokus pada akurasi, tetapi harus mempertimbangkan validitas data dan potensi bias yang mempengaruhi reliabilitas hasil.

Depresi pada mahasiswa bersifat multifaktoral dan dipengaruhi oleh interaksi antara faktor gaya hidup, demografi, dan psikologis. Faktor gaya hidup seperti pola tidur, aktivitas fisik, dan manajemen stres berkontribusi langsung terhadap kondisi kesehatan mental [8], [12], sementara faktor demografis dan psikologis turut menentukan tingkat kerentanan individu terhadap depresi [7], [9]. Namun, sebagian besar penelitian terdahulu masih mengkaji faktor-faktor tersebut secara terpisah, sehingga hubungan antar variabel belum terintegrasi secara komprehensif dalam menjelaskan tingkat depresi.

Selain itu, meskipun banyak penelitian melaporkan performa model yang tinggi, evaluasi terhadap validitas model dan potensi bias masih terbatas. Perbandingan antar algoritma dalam konteks dataset yang sama juga belum dilakukan secara komprehensif, sehingga keunggulan relatif suatu metode belum dapat dipastikan. Beberapa studi menunjukkan bahwa performa algoritma seperti *Support Vector Machine, Decision Tree, Random Forest,* dan *K-Nearest Neighbors* sangat bergantung pada karakteristik data yang digunakan [6], [10], [11]. Kondisi ini menunjukkan bahwa performa tinggi belum tentu mencerminkan model yang robust dan dapat diandalkan dalam implementasi nyata.

Berdasarkan keterbatasan tersebut, terdapat celah penelitian yang signifikan, yaitu belum adanya pendekatan klasifikasi tingkat depresi yang secara simultan: (1) mengintegrasikan faktor gaya hidup, demografi, dan psikologis dalam satu kerangka model yang komprehensif, (2) mempertimbangkan potensi *overfitting*, *data leakage*, dan *bias dataset* dalam proses evaluasi, serta (3) melakukan validasi performa melalui perbandingan dengan algoritma *machine learning* lain sebagai *baseline*. Oleh karena itu, penelitian ini mengusulkan penggunaan Random Forest sebagai model utama, serta Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) sebagai model pembanding. SVM efektif dalam menangani data berdimensi tinggi melalui pembentukan *hyperplane* optimal, sedangkan KNN mampu menangkap pola lokal pada data, sehingga keduanya umum digunakan dalam klasifikasi kesehatan mental [10], [15].

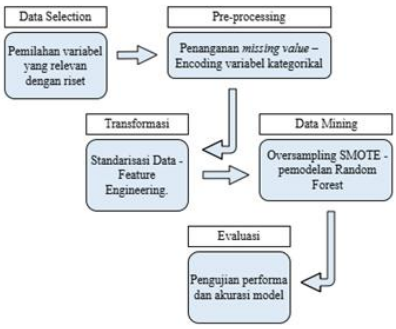
Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model klasifikasi tingkat depresi berbasis integrasi faktor gaya hidup, demografi, dan psikologis. Secara akademis, penelitian ini berkontribusi dalam pengembangan model *machine learning* yang lebih robust dengan mempertimbangkan aspek validitas dan potensi bias data. Secara praktis, model yang dihasilkan diharapkan dapat menjadi dasar dalam pengembangan sistem deteksi dini depresi yang lebih akurat dan aplikatif pada populasi mahasiswa dalam konteks pendidikan tinggi.

2. METODE PENELITIAN

2.1 Pendekatan Penelitian

Penelitian ini menganalisis kumpulan data Klasifikasi Kesehatan Mental dari repositori

Mendeley Data dengan menggunakan pendekatan *Knowledge Discovery in Databases* (KDD). Algoritma *Random Forest* dipilih karena kemampuannya dalam menangani kumpulan data berdimensi tinggi, mengurangi *overfitting* melalui teknik *ensembel*, serta mengidentifikasi hubungan non linier antar variabel yang digunakan dalam model prediksi.



Gambar 1. Metodologi Penelitian

2.2 Tahapan Metode KDD

Proses *Knowledge Discovery in Database* (KDD) dalam penelitian ini dilakukan melalui lima tahapan utama, yaitu :

2.2.1 Seleksi Data (*Data Selection*)

Pada tahap ini, dipilihlah variabel-variabel yang berkaitan dengan kondisi kesehatan mental dan faktor gaya hidup individu. Tujuan dari proses seleksi ini adalah untuk mengurangi gangguan data dan memastikan bahwa atribut-atribut yang digunakan memiliki pengaruh terhadap klasifikasi tingkat depresi. Pendekatan *Chi-Square*, sebuah teknik statistik yang digunakan untuk menilai tingkat hubungan antara faktor prediktor dan variabel target (Jenis Depresi), digunakan untuk memilih fitur-fitur untuk penelitian ini. Model Random Forest kemudian dibangun menggunakan hasil dari pemilihan fitur ini untuk memperkirakan tingkat keparahan depresi [16].

2.2.2 Pra-Pemrosesan (*Pre-processing*)

Pada tahap ini, diterapkan berbagai teknik pengolahan data untuk meningkatkan kualitas data sebelum digunakan dalam model *machine learning*.
a. Penanganan *missing values*: Data yang tidak lengkap diisi menggunakan metode modus (*most frequent value imputation*) [16].
b. Encoding Data Kategorikal: Variabel seperti jenis kelamin dan negara dikonversi ke dalam format numerik menggunakan label *Encoding* [16].

2.2.3 Transformasi Data (*Transformation Data*)

Pada tahap transformasi data, dataset dibagi menjadi dua bagian, yaitu data latih *(train)* sebesar 70% dan data uji *(test)* sebesar 30%. Proporsi ini dipilih untuk menjaga keseimbangan antara kebutuhan data pelatihan yang cukup besar serta kemampuan evaluasi model terhadap data yang belum pernah digunakan sebelumnya [16].

2.2.4 Penambangan Data (*Data Mining*)

a. *Oversampling* dengan SMOTE: Metode *Synthetic Minority Over-sampling* (SMOTE) digunakan untuk mengatasi ketidakseimbangan distribusi kelas dalam kumpulan data. Teknik ini bekerja dengan menghasilkan sampel sintetis pada kelompok minoritas sehingga distribusi data menjadi lebih seimbang. Dengan cara ini, model dapat mempelajari pola pada setiap kelas secara lebih representatif dan mengurangi bias terhadap kelas mayoritas [17].
b. Model *Random Forest*: Tahap ini merupakan inti dari penelitian, di mana model *Random Forest Classifier* dilatih untuk melakukan klasifikasi tingkat depresi berdasarkan faktor gaya hidup, atribut demografi, serta kondisi psikologis responden. Model ini menggunakan parameter *class_weight = 'balanced'* untuk menangani ketidakseimbangan data [17].

2.2.5 Evaluasi (*Evaluation*)

Performa model dievaluasi menggunakan beberapa metrik berikut:
a. Akurasi *(Accuracy)*: metrik ini merepresentasikan persentase total prediksi yang tepat (positif dan negatif) dibandingkan dengan keseluruhan jumlah data observasi [16].

$$Accuracy = \frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$$

b. Presisi (*Precision*): menilai sejauh mana model mampu memberikan prediksi yang akurat pada individu yang benar-benar mengalami gangguan mental di antara seluruh hasil prediksi positif [16].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

c. Sensitivitas (*Recall*): menunjukkan sejauh mana model mampu mengenali kembali seluruh sampel yang secara aktual berada di kelas positif [16].

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

d. F1-Score: merupakan nilai rata-rata harmonik antara precision dan recall yang memberikan gambaran keseimbangan performa model [16].

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

e. AUC-ROC (*Area Under the Curve - Receiver Operating Characteristic*): mengukur kemampuan diskriminasi model dalam membedakan antara kelas positif dan negatif. Nilai AUC yang mendekati angka 1 mengindikasikan performa model yang sangat optimal [16].

Selain evaluasi berbasis metrik, penelitian ini juga menerapkan *Stratified K-Fold Cross-Validation* dengan 5 fold untuk menguji kestabilan dan kemampuan generalisasi model. Pendekatan ini dipilih karena mampu mempertahankan proporsi distribusi kelas pada setiap fold, sehingga memberikan evaluasi yang lebih representatif terhadap kondisi dataset yang tidak seimbang [18]. Nilai performa yang dilaporkan merupakan rata-rata hasil dari proses *cross-validation*, sehingga mencerminkan kinerja model yang lebih stabil dan konsisten pada berbagai pembagian data.

3. HASIL DAN PEMBAHASAN

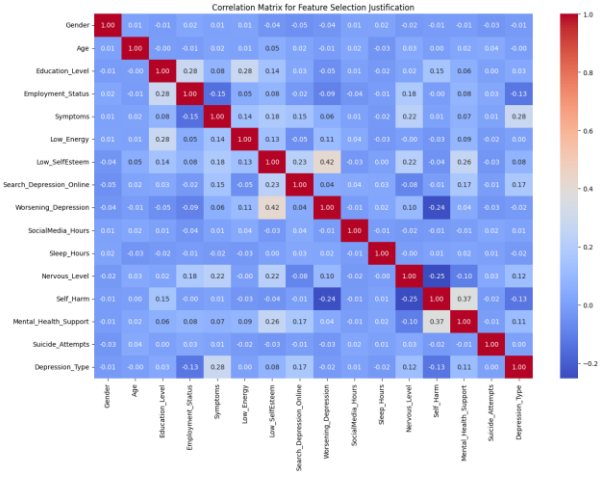
3.1 Seleksi Data (*Data Selection*)

Penelitian ini menggunakan dataset *Mental Health Classification* yang bersumber dari repositori Mendeley Data. Dataset berisi berbagai atribut terkait kondisi kesehatan mental dan faktor gaya hidup individu. Mencakup variabel gaya hidup, demografi, serta kondisi psikologis. Dataset awal terdiri dari 1998 data dengan 21 fitur.

Tabel 1. Chi-Square

Rank	Feature	Chi-Square Score
1	*Symptoms*	1053,10
2	*Nervous_Level*	374,67
3	*Mental_Health_Support*	326,39
4	*Search_Depression_Online*	216,06
5	*Self_Harm*	210,83
6	*Low_Energy*	108,19
7	*Employment_Status*	106,51
8	*Worsening_Depression*	103,98
9	*Low_SelfEsteem*	99,61
10	*Education_Level*	51,52
11	*SocialMedia_Hours*	22,56
12	*Sleep_Hours*	21,65
13	*Suicide_Attempts*	16,56
14	*Gender*	4,98
15	*Age*	2,76

Untuk menghindari bias subjektivitas, seleksi fitur tidak dilakukan secara arbitrer, melainkan divalidasi melalui uji dependensi statistik *Chi-Square*. Proses ini menghasilkan 16 variabel, terdiri dari 15 fitur prediktor dan 1 variabel target, dengan signifikansi tinggi terhadap *Depression_Type*. Variabel *Symptoms* memiliki nilai *importance* tertinggi (1053,09), diikuti *Nervous_Level* (374,67) dan *Mental_Health_Support* (326,38).



Gambar 2. *Correlation Matrix*

Analisis *Correlation Matrix* menunjukkan bahwa korelasi antara *Symptoms* dan target bersifat moderat (0,28), sehingga *Symptoms* tetap memberikan informasi esensial tanpa menjadi duplikasi label, dan integritas model tetap terjaga.

3.2 Pra-Pemrosesan (*Pre-processing*)

Pada tahap *preprocessing*, tidak ditemukan *missing value* (dataset lengkap). Namun, sebagai mitigasi, *imputation* tetap disiapkan menggunakan modus untuk data kategorikal dan median untuk numerik guna menjaga kesiapan pada data *real-time*. Selanjutnya, dilakukan *Label Encoding* untuk mengonversi variabel kategorikal (seperti *Gender*, *Education_Level*, *Employment_Status*, dan *Symptoms*) ke dalam representasi numerik.

3.3 Transformasi Data (*Transformation Data*)

Pada tahap transformasi data, dataset dibagi menjadi data latih (70%) dan data uji (30%) menggunakan metode *train_test_split*. Proses ini menghasilkan 1398 data latih dan 600 data uji dengan 15 atribut fitur. Data latih digunakan dalam proses pelatihan model, sedangkan data uji digunakan untuk menguji kemampuan model dalam melakukan prediksi.

3.4 Penambangan Data (*Data Mining*)

3.4.1 *Oversampling* dengan SMOTE

Diterapkan oversampling menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE). Hal ini disebabkan penelitian yang telah dilakukan menunjukkan hasil yang tidak seimbang. Berdasarkan hasil implementasi SMOTE pada data latih, distribusi kelas target *Depression_Type* menjadi seimbang di mana setiap kelas memiliki 439 data. Dengan distribusi yang lebih merata, model dapat mempelajari pola pada setiap kelas secara lebih optimal dan mengurangi bias terhadap kelas mayoritas.

3.4.2 Model *Random Forest*

Model Klasifikasi *Random Forest* digunakan untuk memprediksi risiko kesehatan mental berdasarkan fitur-fitur yang diberikan setelah proses *oversampling*. Ketidakseimbangan kelas dalam data diatasi dengan menerapkan parameter *class_weight = 'balanced'*.

3.5 Evaluasi dan Komparasi Model

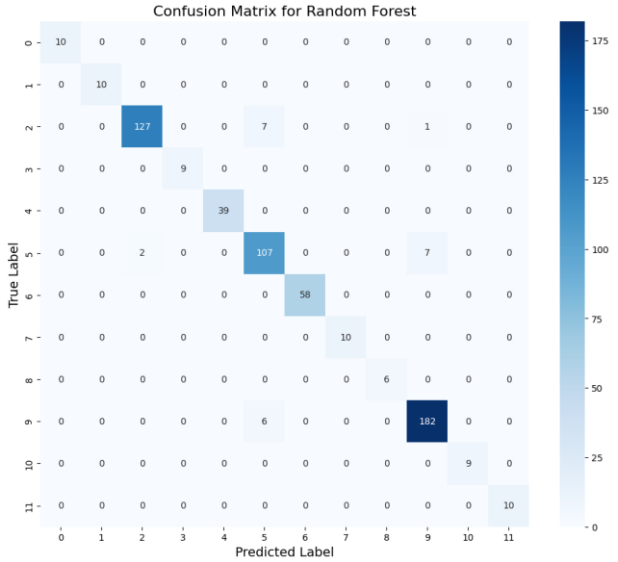
Untuk memastikan objektivitas, performa *Random Forest* dibandingkan dengan *K-Nearest Neighbor (KNN)* dan *Support Vector Machine (SVM)*. Evaluasi tidak hanya menggunakan akurasi, tetapi juga metrik probabilistik seperti *Log Loss* dan *Brier Score* untuk memastikan kalibrasi yang baik serta menghindari *overfitting*.

Tabel 2. Komparasi Metrik Evaluasi Model Klasifikasi

Metrik Evaluasi	Random Forest	K-Nearest Neighbor	Support Vector Machine
Accuracy	**96,17%**	49,83%	44,50%
ROC-AUC (OvR Weighted)	**0,9977**	0,8065	0,8135
Log Loss	**0,1687**	6,5067	1,2792
Brier Score	**0,0163**	0,1235	0,1233

Hasil komparasi pada Tabel 2 menunjukkan dominasi *Random Forest* dengan akurasi 96,17%, jauh melampaui *KNN* (49,83%) dan *SVM* (44,50%). Keunggulan ini logis karena dataset survei kesehatan mental bersifat kompleks, melibatkan data kategorikal ter-*encode* dan *decision boundary* yang non-linear, sehingga lebih sesuai dengan pendekatan *ensemble* berbasis pohon keputusan. Sebaliknya, *KNN* mengalami mengalami penurunan performa akibat *curse of dimensionality*, sedangkan *SVM* menunjukkan keterbatasan dalam memisahkan 12 kelas meski telah menggunakan *SMOTE*. Selain akurasi, *Random Forest* juga mencatat *Log Loss* rendah (0,1687) dan *Brier Score* (0,0163), yang menunjukkan bahwa model tidak hanya akurat, tetapi juga memiliki kalibrasi probabilitas yang baik. Hal ini mengindikasikan bahwa model mampu melakukan generalisasi dengan baik dan tidak sekedar menghafal data latih.

3.6 Analisis Mendalam Matriks Kebingungan *(Confusion Matrix)*

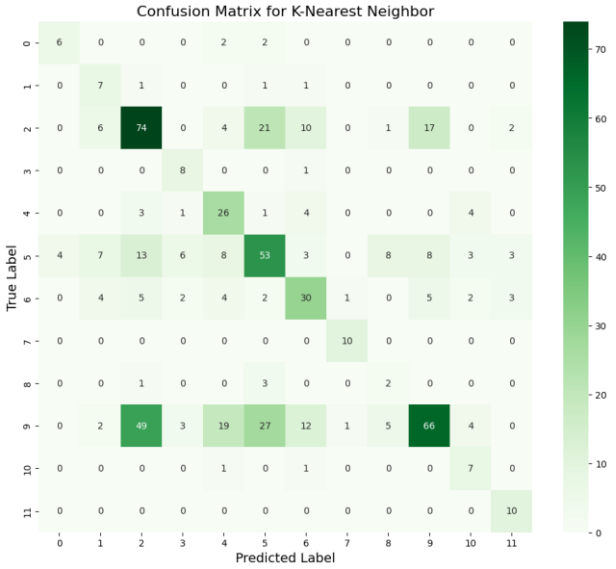


Gambar 3. *Confusion Matrix Random Forest*

Pada model *Random Forest*, *Confusion Matrix* menunjukkan diagonal dominan dengan *precision* dan *recall* mendekati sempurna pada sebagian besar kelas (misalnya kelas 0, 1, 3, 4, 6, 7, 8, 10, dan 11 dengan *recall* 1,00). Namun, terdapat ambiguitas pada kelas 2, 5, dan 9, dengan *misclassification*. Berdasarkan data:

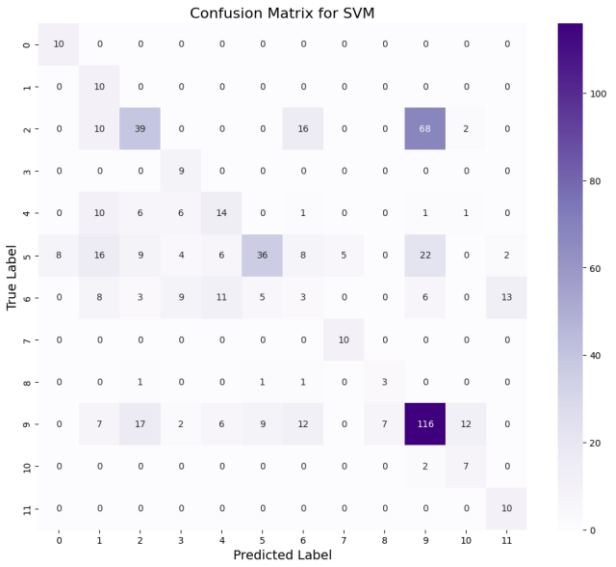
a. Terdapat 7 sampel dari Kelas 2 yang keliru diprediksi sebagai Kelas 5.
b. Terdapat 6 sampel dari Kelas 5 yang keliru diprediksi sebagai Kelas 9, dan sebaliknya (6 sampel Kelas 9 diprediksi sebagai Kelas 5).

Kelas 2, 5, dan 9 merepresentasikan tingkat keparahan depresi yang memiliki karakteristik saling beririsan. Pada data berbasis self-reported, kemiripan gejala di batas antar kelas menimbulkan ambiguitas dalam proses klasifikasi. Dengan demikian, ketidakmampuan model mencapai pemisahan secara sempurna mencerminkan kompleksitas alami data, khususnya dalam membedakan kondisi dengan gejala yang berdekatan.



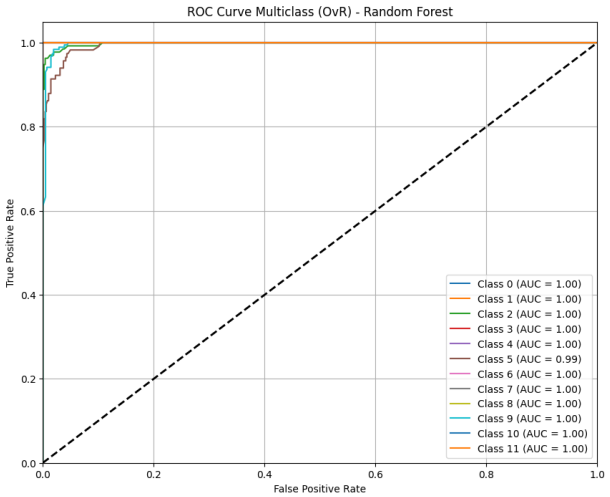
Gambar 4. *Confusion Matrix* KNN

Pada *K-Nearest Neighbor (KNN)*, *Confusion Matrix* menunjukkan distribusi prediksi yang menyebar (*scattered*) akibat keterbatasan metode berbasis jarak dalam menangani data berdimensi tinggi (*curse of dimensionality*).



Gambar 5. *Confusion Matrix* SVM

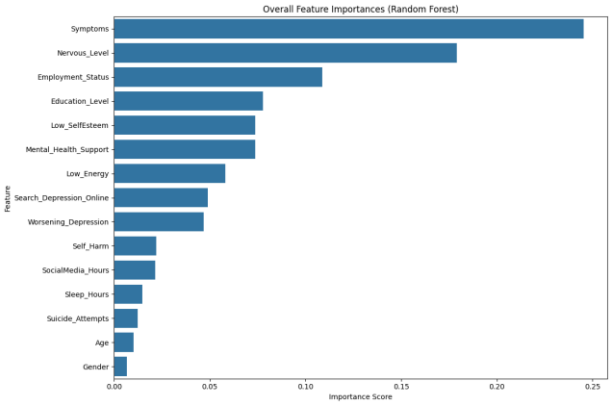
Sementara itu, *Support Vector Machine (SVM)* menunjukkan bias prediksi dengan kecenderungan menebak kelas mayoritas (misalnya kelas 9), yang mengindikasikan kegagalan dalam menangkap pola kompleks dan ketergantungan pada *prior probability*.



Gambar 6. *ROC Curve Multiclass-Random Forest*

Analisis kurva *ROC* (*AUC* per kelas) menguatkan konsistensi *Random Forest* dengan nilai *AUC* di atas 0,99 pada seluruh kelas (termasuk kelas 5 sebesar 0,9947). Hal ini menunjukkan bahwa meskipun terdapat sedikit kesalahan pada batas *threshold*, kemampuan diskriminatif model tetap sangat tinggi.

3.7 Analisis Determinasi Fitur *(Feature Importance)*

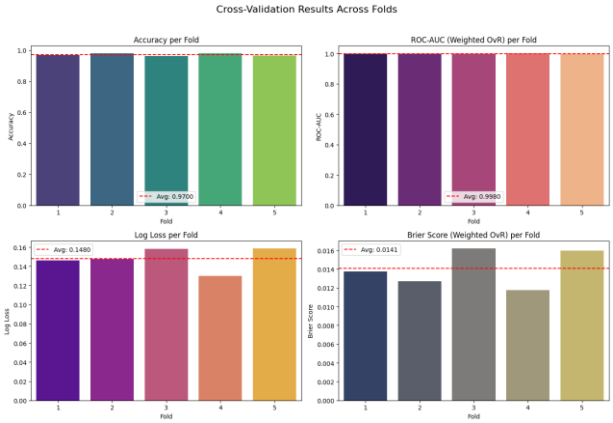


Gambar 7. *Feature Importance* Keseluruhan

Analisis *feature importance* menunjukkan bahwa variabel *Symptoms* dan *Nervous_Level* merupakan prediktor paling dominan dalam model Random Forest. Hal ini sejalan dengan pendekatan klinis yang menempatkan gejala dan tingkat kecemasan sebagai indikator utama kondisi depresi. Selain itu, variabel demografis seperti *Employment_Status* dan *Education_Level* juga memberikan kontribusi signifikan, yang mengindikasikan keterkaitan antara faktor sosial-ekonomi dan kesehatan mental. Dari aspek gaya hidup, *Low_Energy* muncul sebagai faktor penting, yang merefleksikan peran kelelahan fisik dan mental terhadap kerentanan depresi.

Meskipun demikian, dominasi variabel *Symptoms* perlu dikaji secara kritis terkait potensi *label leakage*, mengingat kemungkinan kedekatan konseptual dengan variabel target *Depression_Type* pada data berbasis *self-reported*. Hasil analisis korelasi menunjukkan hubungan pada tingkat moderat, sehingga variabel ini masih dapat dianggap sebagai prediktor yang relevan tanpa secara langsung merepresentasikan label. Selain itu, konsistensi performa model pada proses validasi silang menunjukkan bahwa prediksi tidak hanya bergantung pada satu fitur dominan.

3.8 Validasi Silang *(Cross-Validation)*



Gambar 8. *Stratified K-Fold Cross-Validation*

Kecurigaan terhadap *overfitting* yang sering membayangi akurasi diatas 95% diuji melalui penerapan *Stratified 5-Fold Cross-Validation*. Pengujian ini merotasi partisi data secara sistematis, dan *Random Forest* merespons dengan konsistensi performa yang baik (rata-rata akurasi stabil dengan standar deviasi yang rendah).

Stabilitas kinerja pada setiap lipatan pengujian (*fold*) menunjukkan konsistensi performa model dalam menangkap pola yang relevan pada data. Hal ini mengindikasikan bahwa model tidak hanya menyesuaikan diri terhadap variasi spesifik pada data latih. Temuan tersebut mendukung potensi pemanfaatan model lebih lanjut, termasuk dalam konteks implementasi praktis sebagai instrumen pendeteksi dini di lingkungan institusi pendidikan maupun profesional.

4. KESIMPULAN

Berdasarkan hasil penelitian, model klasifikasi tingkat depresi mahasiswa berhasil dikembangkan menggunakan pendekatan *Knowledge Discovery in Database* (KDD) dengan

algoritma Random Forest. Seleksi fitur menggunakan metode *Chi-Square* menghasilkan 15 fitur prediktor yang relevan terhadap variabel target *Depression_Type*, sementara penerapan SMOTE pada data latih digunakan untuk mengatasi ketidakseimbangan distribusi kelas. Hasil evaluasi menunjukkan bahwa Random Forest memiliki performa terbaik dibandingkan metode *K-Nearest Neighbor* dan *Support Vector Machin*e dengan akurasi sebesar 96,17% dan ROC-AUC 0,9977. Analisis *feature importance* menunjukkan bahwa variabel *Symptoms, Nervous_Level,* dan *Employment_Status* menjadi faktor yang paling berpengaruh dalam klasifikasi tingkat depresi. Hasil *Stratified 5-Fold Cross-Validation* juga menunjukkan performa model yang stabil sehingga memiliki kemampuan generalisasi yang baik. Model yang dihasilkan berpotensi digunakan sebagai dasar pengembangan sistem deteksi dini tingkat depresi pada mahasiswa.

5. SARAN

Penelitian selanjutnya dapat melakukan komparasi model menggunakan algoritma boosting seperti XGBoost, LightGBM, atau CatBoost untuk mengevaluasi potensi peningkatan performa klasifikasi pada data kesehatan mental mahasiswa. Selain itu, optimasi hyperparameter pada model dapat dilakukan secara sistematis menggunakan metode Grid Search atau Bayesian Optimization untuk memperoleh konfigurasi parameter yang lebih optimal. Penelitian selanjutnya juga dapat mempertimbangkan penambahan variabel yang lebih objektif, seperti data aktivitas atau data akademik mahasiswa, sehingga model klasifikasi yang dihasilkan memiliki tingkat validitas dan keandalan yang lebih baik.

DAFTAR PUSTAKA

[1] L. H. Lin, X. Y. Liu, D. Li, Y. K. Zheng, Y. P. Liu, and G. P. Hu, "Predicting 3-year depressive symptoms among middle-aged and older adults in rural China using random forest: insights from the China health and retirement longitudinal study," *BMC Psychol.*, vol. 13, no. 1, Dec. 2025, doi: 10.1186/s40359-025-03513-2.

[2] Y. Dong, H. Wen, C. Lu, J. Li, and Q. Zheng, "Predicting depression risk with machine learning models: identifying familial, personal, and dietary determinants," *BMC Psychiatry*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12888-025-07182-8.

[3] N. Nurdiansyah, F. S. Febriyan, Z. G. D. Amanta, D. A. Saputra, and W. M. Baihaqi, "Analisis Kesehatan Mental untuk Mencegah Gangguan Mental pada Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Random Forest," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 1, pp. 1–9, Nov. 2024, doi: 10.57152/malcom.v5i1.1537.

[4] A. Trifani, S. Tunas Bangsa, A. Perdana, W. Stikom, T. Bangsa, and H. Qurniawan, "Penerapan Data Mining Klasifikasi C4.5 dalam Menentukan Tingkat Stres Mahasiswa Akhir," *Jurnal Riset Rumpun Ilmu Teknik (JURRITEK)*, vol. 1, no. 2, 2022.

[5] K. Yadav and D. T. Bhutia, "Application of machine learning models for predicting depression among older adults with non-communicable diseases in India," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-18053-3.

[6] I. Setiawan, I. Fatah Yasin, Y. Tri Desianti, P. Studi Sistem Dan Teknologi Informasi, F. Sains Dan Teknologi, and A. Surakarta, "Komparasi Kinerja Algoritma Random Forest, Decision Tree, Naïve Bayes, dan KNN dalam Prediksi Tingkat Depresi Mahasiswa Menggunakan Student Depression Dataset," 2025. [Online]. Available: http://creativecommons.org/licences/by/4.0/

[7] A. Aguilar-Latorre, M. J. Serrano-Ripoll, B. Oliván-Blázquez, E. Gervilla, and C. Navarro, "Associations Between Severity of Depression, Lifestyle Patterns, and Personal Factors Related to Health Behavior: Secondary Data Analysis From a Randomized Controlled Trial," *Front. Psychol.*, vol. 13, Mar. 2022, doi: 10.3389/fpsyg.2022.856139.

[8] Z. Zhao, J. Liu, and R. Liu, "Associations of Physical Activity and Sleep Quality with Depression Risk in U.S. Adults: An Interpretable Machine-Learning

Analysis," Nov. 11, 2025. doi: 10.21203/rs.3.rs-7640588/v1.

[9] Y. Wang and S. Xu, "Relationship between artificial intelligence tool usage experience and academic stress among college students: Mediating role of loneliness and moderating role of academic self-efficacy," *Acta Psychol. (Amst).*, vol. 263, Mar. 2026, doi: 10.1016/j.actpsy.2026.106220.

[10] Q. Li and J. Zhou, "A Comparative Analysis of Extreme Gradient Boosting, Decision Tree, Support Vector Machines, and Random Forest Algorithm in Data Analysis of College Students' Psychological Health," *Informatica (Slovenia)*, vol. 49, no. 15, pp. 127–134, 2025, doi: 10.31449/inf.v49i15.7004.

[11] A. Nur, M. Pudjianto, and E. Y. Hidayat, "Perbandingan Prediksi Depresi Mahasiswa dengan Linear Regression, Random Forest, dan Gradient Boosting," *Pendrikan Kidul, Kec. Semarang Tengah*, vol. 207, [Online]. Available: https://doi.org/10.31598

[12] E. Todd *et al.*, "Lifestyle factors and other predictors of common mental disorders in diagnostic machine learning studies: A systematic review," Feb. 01, 2025, *Elsevier Ltd*. doi: 10.1016/j.compbiomed.2024.109521.

[13] A. Chiavegatto Filho, A. F. D. M. Batista, and H. G. dos Santos, "Data Leakage in Health Outcomes Prediction With Machine Learning. Comment on 'Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning,'" *J Med Internet Res*, vol. 23, no. 2, p. e10969, 2021, doi: 10.2196/10969.

[14] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, Oct. 2024, doi: 10.1007/s10462-024-10884-2.

[15] C. M. Sharma and V. M. Chariar, "Diagnosis of mental disorders using machine learning: Literature review and bibliometric mapping from 2012 to 2023," Jun. 30, 2024, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2024.e32548.

[16] M. Maulidah and N. Hidayati, "PREDIKSI KESEHATAN TIDUR DAN GAYA HIDUP MENGGUNAKAN MACHINE LEARNING," *CONTEN : Computer and Network Technology*, vol. 4, no. 1, 2024, [Online]. Available: http://jurnal.bsi.ac.id/index.php/conten

[17] S. S. A. Larasati, E. N. K. Dewi, B. H. Farhansyah, F. A. Bachtiar, and F. Pradana, "Penerapan Decision Tree dan Random Forest dalam Deteksi Tingkat Stres Manusia Berdasarkan Kondisi Tidur," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 5, pp. 1043–1050, Oct. 2024, doi: 10.25126/jtiik.2024117993.

[18] A. Y. Pratama, I. S. Maulana, F. K. Sari, S. D. Tiara, and I. Darmawan, "Prediksi Risiko Depresi pada Mahasiswa Menggunakan Algoritma Random Forest Berdasarkan Data Akademik dan Gaya Hidup," *JSITIK: Jurnal Sistem Informasi dan Teknologi Informasi Komputer*, vol. 4, no. 1, pp. 1–10, Dec. 2025, doi: 10.53624/jsitik.v4i1.696.