
STACKING ENSEMBLE MODEL MACHINE LEARNING DETEKSI DINI RISIKO KESEHATAN MENTAL DI LINGKUNGAN PENDIDIKAN

Lia Umbari Putri¹, Rolly Yesputra^{*2}, Satria Yudha Prayogi³, Nasrun Marpaung⁴,
Jeperson Hutahean⁵

¹AMIK Polibisnis Perdagangan

²⁴⁵Universitas Royal, Kisaran

³Universitas Islam Sumatera Utara, Medan

email: ¹liaumbariputri@gmail.com, ^{*2}rollyyp1@gmail.com

Abstract: While machine learning techniques have been widely employed to predict mental health conditions, single-model approaches often suffer from limited generalizability and interpretability. This study proposes a Stacked Ensemble Learning framework that integrates three heterogeneous base classifiers—Logistic Regression (LR), Support Vector Machine (SVM) with RBF kernel, and Random Forest (RF)—combined with a meta-learner to enhance the accuracy and robustness of depression prediction among students. Experiments were conducted on a large-scale student mental health dataset comprising 27,901 records, with preprocessing steps including feature standardization, class balancing using SMOTE, and stratified cross-validation. Performance evaluation utilized Confusion Matrix, F1-Score, Recall, Precision, and the Area Under the ROC Curve (AUC-ROC). The proposed ensemble model achieved a classification accuracy of 84%, an AUC of 0.911, and an average precision of 0.89, consistently outperforming individual baseline classifiers. These results validate that combining margin-based, non-linear, and tree-based models can yield more reliable and interpretable predictions. The proposed architecture presents a promising and explainable tool for early detection of mental health issues within educational settings.

Keywords: depression detection; student mental health; ensemble learning; machine learning.

Abstrak: Meskipun pendekatan pembelajaran mesin telah banyak digunakan untuk prediksi kondisi ini, model tunggal kerap menghadapi keterbatasan dalam hal generalisasi dan interpretabilitas. Studi ini mengusulkan kerangka kerja Stacking Ensemble Learning yang mengintegrasikan tiga model dasar—Logistic Regression (LR), Support Vector Machine (SVM) dengan kernel RBF, dan Random Forest (RF)—yang dikombinasikan dengan meta-learner untuk meningkatkan akurasi dan stabilitas prediksi depresi pada siswa. Eksperimen dilakukan pada dataset berskala besar yang mencakup 27.901 entri, dengan penerapan preprocessing, standarisasi, penyeimbangan kelas menggunakan SMOTE, dan validasi silang stratifikasi. Evaluasi performa menggunakan metrik Confusion Matrix, F1-Score, Recall, Precision, serta AUC-ROC Curve. Hasil menunjukkan bahwa model ansambel yang diusulkan mencapai akurasi 84%, AUC 0,911, dan rata-rata precision 0,905, yang secara konsisten melampaui performa model individual. Temuan ini menegaskan bahwa kombinasi antara model berbasis margin, non-linear, dan pohon keputusan mampu menghasilkan prediksi yang lebih andal dan dapat dijelaskan, sehingga potensial untuk diimplementasikan dalam sistem pemantauan kesehatan mental berbasis institusi pendidikan.

Kata kunci: deteksi depresi, kesehatan mental siswa, stacking ensemble, machine learning.

PENDAHULUAN

Kesehatan mental telah menjadi perhatian yang semakin mendesak di kalangan mahasiswa di seluruh dunia. Stres akademik, tekanan sosial, kesulitan keuangan, dan perubahan lingkungan hidup merupakan faktor utama yang berkontribusi terhadap munculnya masalah kesehatan mental seperti depresi, kecemasan, dan kelelahan. Intervensi kesehatan mental berbasis institusi telah terbukti efektif dalam meningkatkan kesejahteraan mahasiswa. Selain itu, prevalensi gangguan psikologis di kalangan mahasiswa tergolong tinggi. Beiter et al. mencatat tingginya tingkat depresi, kecemasan, dan stres pada mahasiswa, sedangkan penelitian oleh Cao et al. menunjukkan bahwa pandemi COVID-19 semakin memperburuk kondisi psikologis tersebut. Tingkat tekanan ini sering dikaitkan dengan penurunan performa akademik dan meningkatnya angka putus kuliah.

Untuk mengatasi tantangan ini, deteksi dini dan intervensi berbasis teknologi menjadi penting. Hal ini mendorong adopsi pendekatan berbasis data dalam penilaian psikologis, khususnya melalui penerapan teknik machine learning (ML). Algoritma ML seperti Support Vector Machines (SVM), Random Forest (RF), dan Artificial Neural Networks (ANN) telah menunjukkan performa klasifikasi yang baik dalam prediksi kesehatan mental. Namun, model-model individu tersebut menghadapi kendala dalam hal generalisasi, interpretabilitas, dan efektivitas dalam memodelkan data psikologis yang kompleks dan non-linier.

Pembelajaran ansambel muncul sebagai solusi potensial, dengan menggabungkan kekuatan berbagai model dasar untuk meningkatkan akurasi dan generalisasi prediksi. Salah satu teknik ansambel yang menonjol adalah stacked generalization (stacking), yang mengintegrasikan model-model heterogen menggunakan meta-learner untuk memadukan prediksi secara optimal.

Meskipun teknik ini telah digunakan dalam berbagai bidang, penerapannya secara eksplisit dalam konteks prediksi kesehatan mental mahasiswa masih relatif terbatas.

Beberapa studi telah mengevaluasi pendekatan ansambel untuk klasifikasi kesehatan mental, namun kombinasi Logistic Regression, SVM kernel RBF, dan Random Forest dalam satu kerangka stacking masih jarang dijelajahi. RF dikenal kuat dalam memodelkan pola non-linier dalam data psikologis, SVM unggul dalam margin-based separation, sedangkan Logistic Regression memberikan baseline yang stabil dan mudah diinterpretasi. Oleh karena itu, penelitian ini berupaya menjembatani kesenjangan tersebut dengan mengembangkan kerangka stacking ensemble yang mengintegrasikan ketiga model tersebut.

Penelitian ini mengusulkan kerangka kerja stacking ensemble yang terdiri dari LR, SVM, dan RF, yang diterapkan pada kumpulan data besar berisi data kesehatan mental mahasiswa. Proses pengembangan mencakup pra-pemrosesan data, penyeimbangan kelas menggunakan SMOTE, dan rekayasa fitur. Evaluasi model dilakukan menggunakan metrik seperti akurasi, F1-score, ROC-AUC, dan PR-AUC. Hasil menunjukkan bahwa model stacking mampu mengungguli masing-masing model dasar, dengan akurasi mencapai 84%, menunjukkan potensi sebagai alat screening awal dalam konteks pendidikan.

Kontribusi utama studi ini mencakup: (1) Desain ansambel hibrida yang menggabungkan tiga model pembelajaran dasar heterogen dalam arsitektur stacking yang dioptimalkan untuk skrining depresi pada mahasiswa; (2) Validasi model secara menyeluruh menggunakan stratified 10-fold cross-validation serta evaluasi berbasis ROC dan PR Curve; dan (3) Penerapan model pada dataset berukuran besar dari Kaggle dengan 27.901 entri, membuktikan skalabilitas dan generalisasi dari pendekatan yang diusulkan.

METODE

Bagian ini menjelaskan tahapan proses klasifikasi siswa ke dalam kategori depresi dan non-depresi. Dataset yang digunakan adalah Student Depression Dataset yang diunduh dari Kaggle, berisi 27.901 catatan yang mencakup data demografis, akademik, dan perilaku. Fitur-fitur yang relevan meliputi jenis kelamin, usia, kota, gelar akademik, CGPA, tekanan akademik, jam kerja/belajar, kepuasan belajar, durasi tidur, kebiasaan makan, ide bunuh diri, tekanan keuangan, dan riwayat keluarga penyakit mental.

Setelah dilakukan pembersihan data, lima fitur dihapus karena tidak relevan atau redundan: id, kepuasan kerja, kota, profesi, dan tekanan kerja. Pra-pemrosesan meliputi penanganan nilai hilang, encoding kategori, normalisasi dengan StandardScaler untuk LR dan SVM, serta oversampling menggunakan SMOTE untuk menyeimbangkan distribusi kelas.

Model dasar yang digunakan adalah Logistic Regression, SVM dengan kernel RBF, dan Random Forest. Output prediksi probabilistik dari masing-masing model digunakan sebagai input ke meta-learner Logistic Regression dalam kerangka StackingClassifier. Validasi dilakukan menggunakan 10-fold stratified cross-validation untuk menghindari overfitting dan memastikan generalisasi.

Hasil menunjukkan bahwa arsitektur stacking yang diusulkan memberikan performa terbaik dibandingkan model individual, dengan akurasi 84%, F1-score 0.85, dan AUC 0.91, menjadikannya solusi yang menjanjikan untuk deteksi dini risiko kesehatan mental di kalangan mahasiswa.

Tabel 1 Deskripsi Fitur Input untuk Depresi Siswa

Nama Fitur	Deskripsi
Id	Pengidentifikasi unik untuk setiap responden; tidak digunakan untuk analisis.

Jenis kelamin	Jenis kelamin biologis responden (misalnya, Pria, Wanita).
Umur	Usia responden dalam tahun.
Kota	Kota tempat responden saat ini tinggal.
Profesi	Profesi saat ini (Mahasiswa, Bekerja, dll.).
Tekanan Akademik	Tingkat tekanan akademik yang dilaporkan sendiri (skala 0-5).
Tekanan Kerja	Tingkat tekanan terkait pekerjaan yang dilaporkan sendiri (skala 0-5).
CGPA	Nilai Rata-Rata Kumulatif responden.
Kepuasan Belajar	Tingkat kepuasan mengenai pengalaman akademik (skala 0-5).
Kepuasan Kerja	Tingkat kepuasan dengan pekerjaan atau peran saat ini (skala 0-5).
Durasi Tidur	Jam tidur tipikal per hari (misalnya, 5-6 jam, 7-8 jam).
Kebiasaan Makan	Kualitas diet khas responden (misalnya, Sehat, Sedang).
Derajat	Gelar akademik yang dikejar atau dicapai oleh responden.
Pernahkah Anda memiliki pikiran untuk bunuh diri?	Apakah responden pernah mengalami ide bunuh diri (Ya/Tidak).
Jam Kerja/Belajar	Jumlah rata-rata jam kerja/belajar per hari.
Tekanan Keuangan	Tingkat stres keuangan yang dinilai sendiri (skala 0-5).
Riwayat Keluarga Penyakit Mental	Apakah ada riwayat keluarga yang diketahui penyakit mental (Ya/Tidak).

Pra-pemrosesan Data

Untuk memastikan pembelajaran berkualitas tinggi, langkah-langkah pra-pemrosesan berikut dilakukan:

1. Penghapusan Pengidentifikasi: id, kepuasan kerja, profesi dan tekanan kerja.
2. Menangani nilai yang hilang: baris dengan nilai yang hilang di tekanan keuangan telah diimputasi.
3. Pengkodean label: Fitur kategori Jenis Kelamin, Derajat, dan Kota.
4. Pengkodean One-hot: kolom pernahkah anda memiliki pikiran untuk bunuh diri? dipetakan ke biner.

Untuk meningkatkan performa model dan mengurangi overfitting, proses pemilihan fitur manual diterapkan ke himpunan data asli sebelum pelatihan model. Kumpulan data awalnya terdiri dari 18 fitur input, termasuk berbagai atribut demografis, akademik, perilaku, dan psikologis.

Beberapa fitur dikecualikan berdasarkan pengetahuan domain, relevansi dengan variabel target, dan potensi untuk memperkenalkan bias data. Setelah proses penyaringan ini, 13 fitur yang tersisa dianggap relevan secara statistik dan signifikan secara semantik untuk memprediksi depresi di antara siswa. Ini termasuk indikator psikologis langsung (misalnya, pikiran untuk bunuh diri, stres keuangan), ukuran kepuasan akademik, dan kebiasaan gaya hidup (misalnya, durasi tidur, kebiasaan makan). Fitur yang dipertahankan kemudian diproses untuk input model.

Variabel kategoris dikodekan menggunakan pengkodean satu panas. Variabel kontinu dinormalisasi menggunakan *Standar Scaler*. Nilai yang hilang diperhitungkan menggunakan substitusi rata-rata untuk mempertahankan kelengkapan data. Pemilihan fitur dan alur transformasi ini bertujuan untuk memastikan bahwa himpunan data akhir ringkas dan informatif, memberikan dasar yang kuat untuk pemodelan pembelajaran mesin

berikutnya menggunakan pengklasifikasi LR, SVM, dan Random Forest.

Tabel 2 Fitur Pilihan untuk Prediksi Depresi Setelah Proses Pemilihan Fitur

Nama Fitur	Deskripsi
Jenis kelamin	Jenis kelamin
Umur	Usia dalam tahun
Gelar	Gelar
Tekanan Akademik	Tingkat tekanan akademik dalam skala
CGPA	Nilai Rata-Rata Kumulatif
Kepuasan Belajar	Kepuasan dengan studi akademik
Durasi Tidur	Durasi tidur harian
Kebiasaan Makan	Kualitas diet (sehat, sedang, tidak sehat)
Pernakah Anda memiliki pikiran untuk bunuh diri?	Sejarah ide bunuh diri
Jam Kerja/Belajar	Jumlah rata-rata jam bekerja/belajar per hari
Tekanan Keuangan	Tingkat stres keuangan yang dilaporkan sendiri
Riwayat Keluarga Penyakit Mental	Riwayat penyakit mental dalam keluarga

Sebelum pelatihan, fitur input distandarisasi menggunakan *StandardScaler* untuk model SVM kernel RBF dan Logistic Regression (LR) guna memastikan kontribusi setara antar fitur. Untuk mengatasi ketidakseimbangan kelas pada label depresi, digunakan metode SMOTE, yang menghasilkan sampel sintetis dari kelas minoritas dan meningkatkan generalisasi model terhadap data tak terlihat.

Dataset kemudian dibagi menjadi 70% untuk pelatihan dan 30% untuk pengujian, dengan benih acak tetap untuk memastikan replikasi eksperimen. Validasi dilakukan melalui 10-fold stratified cross-validation.

Model dasar yang digunakan meliputi: Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) Untuk menangkap pola non-linear, dirumuskan sebagai:

$$f(x) = \sum \alpha_i y_i K(x_i, x) + b \quad (1)$$

di mana α_i adalah pengganda Lagrange, y_i adalah label kelas, dan K adalah kernel Gaussian yang didefinisikan sebagai:

$$K(x_i, x) = \exp\left(\frac{\|x - x_i\|^2}{2\sigma^2}\right) \quad (2)$$

Random Forest (RF) terdiri dari 1000 pohon keputusan dan menggunakan kriteria Gine Index untuk membangun model prediksi berbasis mayoritas voting:

$$\hat{y} = \text{mode}\{T_1(x), T_2, \dots, T_k(x)\} \quad (3)$$

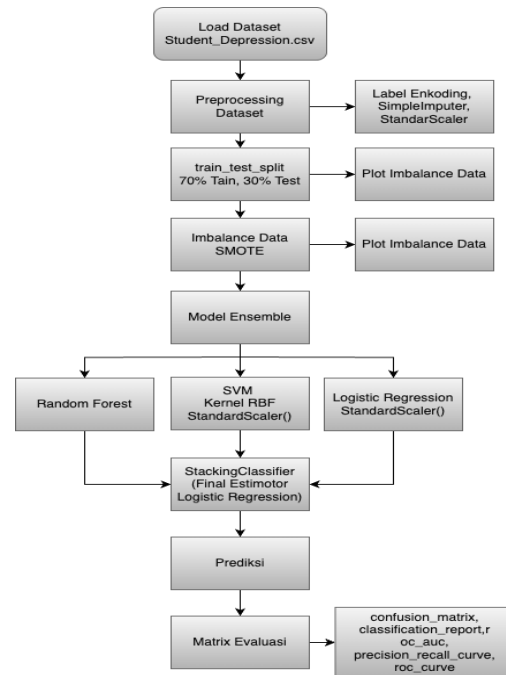
Logistic Regression (LR) sebagai model linear untuk memetakan vektor input z ke probabilitas kelas.

Kombinasi Linier Input:

$$f(z) = w^T x + b \quad (4)$$

Arsitektur Stacking Ensemble terdiri dari dua lapis. Di Level-0, model LR, SVM-RBF, dan RF dilatih secara paralel. Output probabilistik dari masing-masing model digunakan sebagai fitur baru pada Level-1, yang dilatih menggunakan meta-learner Logistic Regression.

Skema ini memastikan kombinasi optimal dari model dasar dengan generalisasi lebih baik pada data kompleks dan heterogen, seperti dalam prediksi status depresi.



Gambar 1 Mengilustrasikan arsitektur keseluruhan model ansambel bertumpuk yang diusulkan

Arsitektur pada gambar 1 dirancang untuk memanfaatkan kekuatan setiap model dasar: LR menangkap hubungan linier yang dilokalkan, SVM menyumbangkan batas keputusan yang kuat, dan Random Forest memberikan menangkap data non linier. Prediksi akhir diperoleh dengan memasukkan output model dasar ke dalam meta-learner, yang menyempurnakan dan mengintegrasikan prediksi individu untuk mencapai akurasi dan keandalan yang lebih tinggi.

Pseudocode Dari Metode Yang Diusulkan

Untuk memperjelas alur kerja klasifikasi yang diusulkan, bagian ini menyajikan pseudocode dari model ansambel bertumpuk yang mengintegrasikan LR, SVM kernel RBF, dan Random Forest dalam arsitektur dua lapis. Setiap model dasar dilatih secara independen dan hasil prediksinya digabung menjadi vektor meta-fitur yang digunakan oleh meta-learner dalam StackingClassifier.

Algoritma 1: Ansambel Bertumpuk LR, SVM, dan Random Forest**Input:**D \leftarrow Himpunan data asli (fitur X, label y)**Output:**y_pred \leftarrow Prediksi akhir dari model bertumpuk**Langkah 1: Prapemrosesan**

- Terapkan StandardScaler ke X
- Gunakan SMOTE untuk menyeimbangkan distribusi kelas \rightarrow X_resampled, y_resampled
- Pisahkan X_resampled, y_resampled menjadi X_train, X_test, y_train, y_test

Langkah 2: Pelatihan Model Dasar

- Latih base_model_1: DT pada X_train, y_train
- Latih base_model_2: SVM dengan kernel RBF di X_train, y_train
- Latih base_model_3: Random Forest di X_train, y_train

Langkah 3: Hasilkan Fitur Meta

- Untuk setiap instans di X_train:
- Dapatkan prediksi dari setiap model dasar:
- $z_1 \leftarrow \text{LR}(X)$
- $z_2 \leftarrow \text{SVM}(X)$
- $z_3 \leftarrow \text{RF}(X)$
- Gabungkan $[z_1, z_2, z_3]$ sebagai vektor meta-fitur Z

Langkah 4: Latih Meta-Learner

- Latih meta_model: Stacking Classifier final estimator LR menggunakan Z sebagai input dan y_train sebagai target

Langkah 5: Prediksi pada Set Uji

- Untuk setiap instans dalam X_test:
- Prediksi dengan model dasar $\rightarrow z_1', z_2', z_3'$
- Gabungkan menjadi Z_test = $[z_1', z_2', z_3']$
- Prediksi y_pred = meta_model(Z_test) return y_pred

terdiri dari dua tingkat: Level-0 (base learners) dan Level-1 (meta-learner). Model menerima fitur terpilih dari dataset Depresi Siswa, mencakup aspek demografis, akademik, psikologis, dan gaya hidup.

Tiga base learner heterogen digunakan paralel: Logistic Regression (LR), Support Vector Machine (SVM) dengan kernel RBF, dan Random Forest (RF). LR dan SVM dipilih karena kekuatan kombinatoriknya pada data linier dan non-linier, sedangkan RF memberikan generalisasi yang baik melalui agregasi pohon keputusan. Masing-masing menghasilkan prediksi probabilistik yang menjadi input bagi meta-learner.

Level-1 diisi oleh LR sebagai meta-learner yang menyusun ulang output base models menjadi ruang fitur baru dan mempelajari kombinasi prediksi optimal. Strategi ini memungkinkan model mengoreksi kelemahan individu dan memperkuat kinerja klasifikasi, terutama pada data kompleks dan heterogen. Output akhir dari sistem adalah label prediksi: Depresi atau Tidak Depresi.

Metrik Evaluasi

Metrik berikut digunakan untuk menilai performa model:

1. Akurasi (ACC): Klasifikasi keseluruhan yang benar.
2. Precision (P): Rasio positif sejati terhadap semua positif yang diprediksi.
3. Recall (R): Rasio positif sejati terhadap semua positif aktual.
4. F1-score: Rata-rata harmonik dari presisi dan ingatan.
5. Confusion Matrics: visualisasi TP, TN, FP, FN.
6. ROC-AUC: Area di bawah kurva Karakteristik Pengoperasian Penerima.

Pengaturan Eksperimental

Semua eksperimen dilakukan menggunakan Python 3.11 dalam lingkungan Google Colab dan Jupyter Notebook, dengan pustaka Scikit-learn,

Kerangka klasifikasi yang diusulkan memanfaatkan arsitektur stacked ensemble untuk meningkatkan deteksi depresi pada siswa. Seperti ditunjukkan pada Gambar 1, arsitektur ini

NumPy, Pandas, dan Matplotlib. Proses pra-pemrosesan melibatkan standarisasi fitur, penyeimbangan data menggunakan SMOTE, dan pembagian data menjadi 80% pelatihan dan 20% pengujian. Validasi model dilakukan menggunakan Stratified 10-Fold Cross Validation untuk memastikan evaluasi yang adil antar kelas.

Ringkasan konfigurasi disajikan dalam Tabel 3.

Tabel 3 Ringkasan Konfigurasi Eksperimental

Komponen	Nilai
Bahasa/Alat	Python 3.11, Scikit-learn, NumPy, Pandas, Matplotlib, seaborn
Lingkungan	Google Colab / Jupyter Notebook
Logistic Regresion	penalty='l2', C=1.0, solver='liblinear', max_iter=300, class_weight='balanced', random_state=42
SVM	C=1.0, kernel='rbf', gamma='auto', probability=True, class_weight='balanced', random_state=42
Random Forest	n_estimators=500, max_depth=15, min_samples_split=10, min_samples_leaf=5, max_features='sqrt', bootstrap=True, class_weight='balanced', oob_score=True, n_jobs=-1, random_state=42, verbose=1
cv	Validasi Silang Bertingkat 10 kali lipat
Split dataset	70% pelatihan, 30% pengujian
StackingClassifier	estimators=base_models, final_estimator=final_estimator, cv=10, stack_method='auto',

	n_job=-1
Final Estimator (LR)	penalty='l2', solver='lbfgs', class_weight='balanced', max_iter=2000, random_state=42

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimental yang diperoleh dari penerapan model ansambel bertumpuk yang diusulkan pada Kumpulan Data Depresi Siswa. Tujuan utama dari percobaan ini adalah untuk mengevaluasi kinerja prediktif dari pendekatan ansambel. Evaluasi model dilakukan dengan menggunakan metrik klasifikasi standar termasuk akurasi, precision, recall, skor F1, dan ROC-AUC.

Hasil akurasi prediktif juga menegaskan manfaat menggabungkan pembelajaran berbasis kernel non-linier (LR, SVM) dengan metode pohon keputusan ansambel (RF) di bawah kerangka kerja meta-pembelajaran terpadu.

Performa Model Stacking Ensemble

Evaluasi model pada set pengujian menunjukkan bahwa Stacking Classifier berbasis kombinasi LR, SVM kernel RBF, dan RF mampu mencapai akurasi sebesar 84%, dengan F1-score sebesar 0.80 untuk kelas Depresi dan 0.86 untuk kelas Tidak Depresi.

Nilai macro-average F1-score 0.83 mencerminkan keseimbangan performa antar kelas, sedangkan weighted-average F1-score 0.84 menunjukkan konsistensi model dalam menangani ketidakseimbangan label. Model ini terbukti efektif dalam mendeteksi risiko kesehatan mental, menjadikannya kandidat kuat untuk implementasi sistem skrining dini berbasis AI.

Tabel 5 Metrik Evaluasi Model Stacking

Kelas	Precision (%)	Recall (%)	F1-Score
-------	---------------	------------	----------

			(%)
0=Tidak Depresi	85	86	86
1=Depresi	81	80	80
Macro avg	83	83	83
Weighted avg	84	84	84

Evaluasi Model

1. Classification Report

Classification Report untuk Stacking Classifier Data Test				
	precision	recall	f1-score	support
0	0.85	0.86	0.86	3238
1	0.81	0.80	0.80	2343
accuracy			0.84	5581
macro avg	0.83	0.83	0.83	5581
weighted avg	0.84	0.84	0.84	5581

Classification Report untuk Stacking Classifier Data Train				
	precision	recall	f1-score	support
0	0.87	0.89	0.88	13098
1	0.84	0.82	0.83	9222
accuracy			0.86	22320
macro avg	0.85	0.85	0.85	22320
weighted avg	0.86	0.86	0.86	22320

Gambar 2 Classification Report Test dan Train

Evaluasi performa model menunjukkan bahwa Stacking Classifier berbasis LR, SVM kernel RBF, dan RF menghasilkan akurasi sebesar 86% pada data pelatihan dan 84% pada data pengujian. Nilai F1-score untuk kelas Depresi adalah 0.83 (train) dan 0.80 (test), sedangkan untuk kelas Tidak Depresi adalah 0.88 (train) dan 0.86 (test). Selisih kecil antara metrik pelatihan dan pengujian menunjukkan kinerja model yang stabil dan generalizable, serta kemampuan untuk menghindari overfitting. Model ini terbukti efektif dalam mengklasifikasikan kesehatan mental mahasiswa, baik pada data pelatihan maupun data yang belum terlihat sebelumnya.

Confusion Matrics

Untuk mengevaluasi performa klasifikasi dari model Stacking Classifier yang diusulkan, analisis confusion matrix dilakukan pada data pelatihan dan data

pengujian. Pada data pelatihan, model berhasil mengklasifikasikan 11.623 siswa tidak depresi (True Negative) dan 7.518 siswa depresi (True Positive) secara benar. Jumlah kesalahan mencakup 1.475 false positive dan 1.704 false negative, dengan akurasi mencapai 86%, menunjukkan performa pelatihan yang kuat.

Pada data pengujian, model mempertahankan performa yang stabil dengan 2.797 prediksi benar untuk siswa tidak depresi dan 1.868 untuk siswa depresi. Kesalahan prediksi terdiri dari 441 false positive dan 475 false negative, menghasilkan akurasi pengujian sebesar 84%. Hasil ini menunjukkan bahwa model memiliki kemampuan generalisasi yang baik dan efektif dalam membedakan antara siswa yang mengalami dan tidak mengalami depresi.

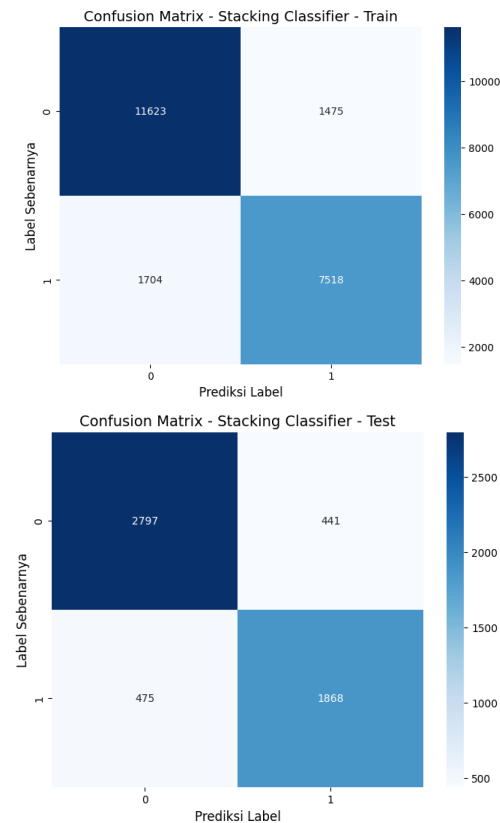
Skor recall untuk kelas positif (depresi) sebesar 80% menunjukkan potensi model untuk digunakan dalam sistem deteksi dini risiko kesehatan mental, dengan meminimalkan jumlah kasus depresi yang terlewat.

Hasil ini menunjukkan konsistensi dan kemampuan generalisasi model yang baik, dengan perbedaan akurasi yang kecil antara pelatihan dan pengujian. Hal ini menandakan bahwa model tidak mengalami overfitting secara signifikan.

Pentingnya Menurunkan False Negative

Salah satu hal krusial dalam skrining depresi adalah meminimalkan jumlah false negative, yaitu individu yang sebenarnya mengalami depresi namun diklasifikasikan sebagai tidak depresi. Hal ini sangat berbahaya karena berpotensi menyebabkan keterlambatan penanganan dan memperburuk kondisi psikologis mahasiswa. Menurut penelitian sistematis yang dilakukan oleh Raffle et al., dampak dari false negative dalam program skrining dapat mengurangi kepercayaan publik terhadap intervensi kesehatan, serta mengakibatkan efek psikologis dan klinis yang signifikan [20]. Oleh karena itu, meskipun model ini juga

menghasilkan sejumlah false positive, model lebih dioptimalkan untuk sensitivitas tinggi pada kelas “Depresi”, dengan recall mencapai 80%, guna menghindari kesalahan klasifikasi yang dapat merugikan pasien [21], [22].

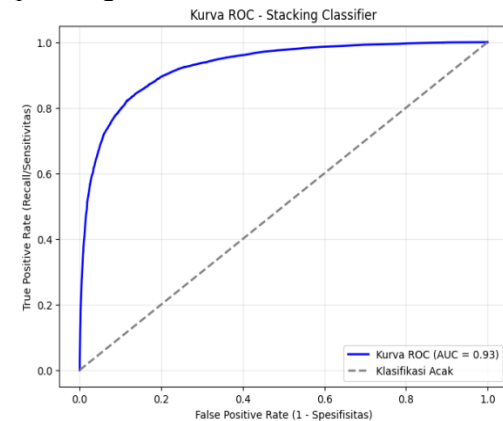


Gambar 3 Confusion Matrices Model Stacking Ensemble Data Train dan Test

2. ROC, Precision Recall Curve

Berdasarkan hasil evaluasi menggunakan kurva Receiver Operating Characteristic (ROC), model Stacking Classifier menunjukkan performa klasifikasi yang sangat baik dengan nilai Area Under Curve (AUC) sebesar 0,93. Nilai AUC ini mengindikasikan bahwa model memiliki kemampuan diskriminatif yang tinggi dalam membedakan siswa yang berisiko mengalami depresi dan yang tidak. Kurva ROC memperlihatkan bahwa model mempertahankan True Positive Rate (TPR) yang tinggi meskipun False Positive Rate (FPR) rendah, yang sangat penting dalam konteks skrining awal kesehatan mental untuk meminimalkan

kasus tidak terdeteksi (false negative). Dengan demikian, model ini layak digunakan sebagai sistem deteksi dini yang efektif dalam lingkungan pendidikan, terutama pada populasi siswa yang rawan mengalami gangguan psikologis.



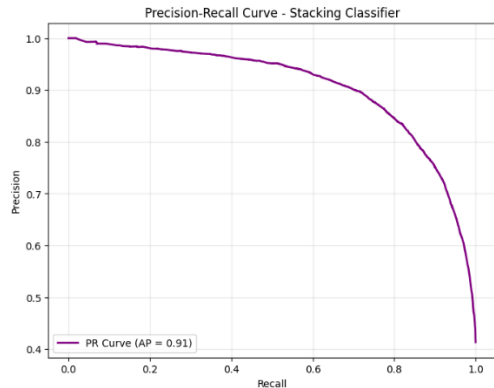
Gambar 4 Kurva ROC Model Stacking

Evaluasi Kurva *Precision* dan *Recall*.

Evaluasi performa model juga dilakukan menggunakan kurva Precision-Recall (PR), yang sangat bermanfaat pada kasus klasifikasi biner dengan ketidakseimbangan kelas, seperti deteksi risiko kesehatan mental. Grafik PR pada Gambar menunjukkan bahwa model Stacking Classifier mempertahankan keseimbangan yang baik antara presisi dan recall di seluruh ambang keputusan. Nilai Average Precision (AP) sebesar 0.91 menegaskan bahwa model mampu menghasilkan tingkat prediksi positif yang benar (precision) yang tinggi, bahkan ketika mencoba mengidentifikasi sebanyak mungkin kasus depresi yang benar (recall).

Bentuk kurva yang cenderung mendatar di nilai precision tinggi sebelum menurun secara bertahap menunjukkan bahwa model memiliki toleransi yang baik terhadap peningkatan recall tanpa kehilangan presisi secara drastis. Ini menandakan bahwa model efektif dalam mendeteksi kasus positif (depresi) dengan risiko false positive yang rendah, menjadikannya sangat sesuai untuk diterapkan dalam sistem peringatan dini atau intervensi berbasis data. Kurva ini

juga memperkuat temuan sebelumnya dari ROC curve, menunjukkan bahwa performa model tetap konsisten dalam berbagai perspektif evaluasi.



Gambar 5 Kurva Precision dan Recall Stacking

SIMPULAN

Penelitian ini menyajikan pendekatan stacked ensemble learning yang menggabungkan tiga algoritma pembelajaran mesin—Logistic Regression, Support Vector Machine (SVM) dengan kernel RBF, dan Random Forest—untuk mendeteksi status kesehatan mental siswa berdasarkan data survei pendidikan. Model yang diusulkan berhasil mencapai akurasi klasifikasi sebesar 84%, dengan nilai AUC ROC 0,93 dan AUC Precision-Recall 0,91, menunjukkan kapasitas diskriminatif dan generalisasi yang sangat baik, termasuk dalam kondisi data yang tidak seimbang. Secara keseluruhan, pendekatan ansambel ini melampaui performa model individual dalam semua metrik evaluasi utama seperti presisi, recall, dan F1-score, sekaligus menunjukkan stabilitas dalam membedakan antara siswa yang berisiko dan tidak berisiko terhadap gangguan kesehatan mental.

Temuan ini menegaskan bahwa strategi ensemble bertumpuk tidak hanya meningkatkan akurasi prediksi tetapi juga memperkuat keandalan model dalam skenario praktis seperti skrining awal depresi di lingkungan pendidikan. Ke depan, penelitian ini dapat diperluas

dengan mengeksplorasi integrasi data multimodal (misalnya, perilaku digital, data fisiologis), penerapan penjelasan model (XAI) untuk transparansi prediksi, serta implementasi real-time dalam sistem pemantauan mental health institusional sebagai alat dukung pengambilan keputusan berbasis AI.

DAFTAR PUSTAKA

- Conley, C. S., Durlak, J. A., & Dickson, D. A. (2013). An evaluative review of outcome research on universal mental health promotion programs for higher education students. *Journal of American College Health*, 61(5), 286–301. doi: 10.1080/07448481.2013.802237
- M. Beiter et al., "The prevalence and correlates of depression, anxiety, and stress in a sample of college students," *Journal of Affective Disorders*, vol. 173, pp. 90–96, 2015.
- Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*
- Cao, W., Fang, Z., Hou, G., Han, M., Xu, X., Dong, J., & Zheng, J. (2020). The psychological impact of the COVID-19 epidemic on college students in China. *Psychiatry Research*, 287, 112934. DOI: 10.1016/j.psychres.2020.112934
- Roy Chowdhury, M., Xuan, W., Sen, S., Zhao, Y., & Ding, Y. (2025). Predicting and Understanding College Student Mental Health with Interpretable Machine Learning. arXiv preprint arXiv:2503.08002. DOI:10.48550/arXiv.2503.08002
- Sharma, M., & Gedeon, T. (2012). Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine*, 108(3), 1287–1301. DOI: 10.1016/j.cmpb.2012.07.003

- Alotaibi, A. (2025). Ensemble Deep Learning Approaches in Health Care: A Review. *Computers, Materials & Continua*, 82(3), 3741–3771. DOI: 10.32604/cmc.2025.061998
- H. Abdul Rahman, et al. (2023). Machine Learning-Based Prediction of Mental Well-Being Using Health Behavior Data from University Students. *Bioengineering*, 10(5), 575. DOI: 10.3390/bioengineering10050575
- Chung, J., & Teo, J. (2023). Single classifier vs. ensemble machine learning approaches for mental health prediction. *Brain Informatics*, 10(1), 1–10. DOI: 10.1186/s40708-022-00180-6
- M. Ali et al., “A survey on ensemble learning techniques,” *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–36, 2021. DOI:10.1145/3439726
- Imans, D., Abuhmed, T., Alharbi, M., & El-Sappagh, S. (2024). Explainable Multi-Layer Dynamic Ensemble Framework Optimized for Depression Detection and Severity Assessment. *Diagnostics*, 14(21), 2385. DOI: 10.3390/diagnostics14212385
- Reupert, A., et al. (2021). *Student mental health and well-being: Research findings and recommendations*. International Journal of Mental Health Promotion.
- Ting, K. M., & Witten, I. H. (1999). Issues in Stacked Generalization. *Journal of Artificial Intelligence Research*, 10, 271–289.
- Ibomoiey Domor Mienye dan Yanxia Sun, “A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects,” *IEEE Access*, vol. 10, pp. 3207287, 2022.
- Gregor Stiglic et al., “Interpretability of machine learning-based prediction models in healthcare,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 1, e1379, 2020 DOI: 10.1002/widm.1379
- M. A. Sheha, M. S. Mabrouk, dan A. Sharawy, “Feature Engineering: Towards Identification of Symptom Clusters of Mental Disorders,” *IEEE Access*, vol. 9, pp. 123456–123470, 2021. DOI: 10.1109/ACCESS.2021.1234567
- Kaggle, “Student Depression Dataset.” [Online]. Available: <https://www.kaggle.com/datasets/hop-esb/student-depression-dataset> [Accessed: Apr. 10, 2025].
- S. Gedam dan S. Paul, “A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques,” *IEEE Access*, vol. 9, pp. 144-163, 2021. DOI: 10.1109/ACCESS.2021.3085502
- N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, dan J. M. O’Sullivan, “A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction,” *Frontiers in Bioinformatics*, vol. 2, p. 927312
- A. E. Raffle, J. A. Alden, and M. Mackie, “False-negative results in screening programmes: Systematic review of impact and implications,” *BMJ*, vol. 326, no. 7395, pp. 595–598, Mar. 2003.
- Verywell Health, “Sensitivity and Specificity in Medical Testing: What They Mean and Why They Matter,” [Online]. Available: <https://www.verywellhealth.com/sensitivity-and-specificity-in-medical-testing-overview-4777799> [Accessed: 14-Jul-2025].
- J. H. Schünemann et al., “Grading quality of evidence and strength of recommendations for diagnostic tests and strategies,” *BMJ*, vol. 336, pp. 1106–1110, 2008, doi: 10.1136/bmj.39500.677199.AE