Prediksi Tingkat Stres Berdasarkan Pola Hidup Menggunakan Machine Learning

Amanda Istiqomatul Anissa[1], Anita Qoiriah[2]

[1,2] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Negeri Surabaya
[1]amanda.21057@mhs.unesa.ac.id
[2]anitaqoiriah@unesa.ac.id

***Abstrak*— Stres merupakan kondisi psikologis yang dapat mengakibatkan dampak yang serius terhadap kesehatan mental dan fisik individu. Penelitian ini bertujuan untuk mengembangkan model prediksi tingkat stres berdasarkan pola hidup individu menggunakan pendekatan *Machine Learning* dengan algoritma *Random Forest*. Proses pengembangan model meliputi pengumpulan *dataset* pola hidup individu, *preprocessing* data, pembagian data, Penyeimbangan data dengan SMOTE serta pelatihan dan evaluasi model *Random Forest*. Data pada penelitian ini diperoleh melalui kuesioner mencakup variabel-variabel pola hidup seperti kualitas tidur, aktivitas fisik, konsumsi kafein, kebiasaan merokok, dan konsumsi alkohol serta menggunakan hasil dari pengukuruan *Skala Perceived Stress Scale* (PSS-10) dengan total 204 data. Evaluasi model dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*, menghasilkan akurasi tertinggi 83% dengan menggunakan *Random Forest* yang diuji menggunakan teknik Stratified K-Fold Cross Validation. Dengan demikian, penelitian ini diharapkan dapat menjadi dasar bagi sistem pendukung keputusan dalam upaya preventif menjaga kesehatan mental melalui perbaikan pola hidup. Namun, model ini belum dapat dijadikan acuan tunggal dalam penentuan diagnosis stres tanpa mempertimbangkan faktor lain serta validasi klinis lebih lanjut.**

***Kata Kunci*—Stres, pola hidup, *Machine Learning, Random Forest*, PSS-10, prediksi tingkat stres**

I. PENDAHULUAN

Stres adalah respons psikologis dan fisik yang timbul saat seseorang menghadapi tekanan yang melebihi kemampuannya dalam mengatasi situasi. Dalam kehidupan modern, stres menjadi semakin relevan karena ritme hidup yang cepat dan tuntutan yang tinggi sering memicu tekanan mental. Data dari WHO menunjukkan peningkatan global pada gangguan terkait stres dan kesehatan mental [1]. Tak hanya itu, di Indonesia yang mengalami tren kenaikan gangguan emosional menurut Riskesdas, sehingga menegaskan pentingnya perhatian terhadap isu ini [2].

Faktor penyebab stres sangat beragam, baik dari individu maupun lingkungan. Gaya hidup seperti kurangnya aktivitas fisik, pola tidur yang buruk, dan konsumsi kafein, diketahui berkontribusi besar terhadap peningkatan stres [3], [4], [5]. Individu yang aktif secara fisik cenderung memiliki tingkat stres lebih rendah karena meningkatnya hormon seperti dopamin, serotonin, dan endorfin [3], [6]. Sebaliknya, mahasiswa dengan aktivitas fisik rendah berpeluang mendapatkan stres lebih tinggi sebesar 0.286 kali dibandingkan mahasiswa yang melakukan aktivitas sedang-berat [3]. Selain itu, kurang tidur dapat memicu respons emosional negatif dan mengganggu fungsi kognitif [4]. Kafein juga berpengaruh pada peningkatan kortisol, terutama jika dikonsumsi berlebihan, meskipun dalam jumlah sedang bisa memberikan efek positif sementara [5].

Stres juga bisa mendorong perilaku tidak sehat seperti merokok dan konsumsi alkohol [7], [8]. Merokok sering dijadikan pelarian karena nikotin bersifat menenangkan secara sementara karena nikotin dalam rokok memiliki efek psikologis yang dapat sementara mengurangi rasa cemas dan stres [7]. Begitu pula dengan alkohol, yang bisa memberikan efek euforia sesaat akibat stres berat, terutama pada remaja. Sebaliknya, individu dengan stres ringan umumnya tetap mampu fokus pada aktivitas positif dan tidak merasa perlu mengonsumsi alkohol untuk relaksasi [8].

Untuk mengukur persepsi terhadap stres, digunakan instrumen standar bernama Perceived Stress Scale (PSS) yang dikembangkan berdasarkan teori Lazarus dan Folkman. PSS mengevaluasi sejauh mana individu merasa hidupnya tidak terkendali dan sulit diprediksi. Berbeda dari alat ukur stres eksternal, PSS menilai persepsi subjektif terhadap situasi dalam satu bulan terakhir. Terdiri dari 14 item, skala ini menggunakan penilaian ordinal dan mencerminkan tingkat stres secara umum [9].

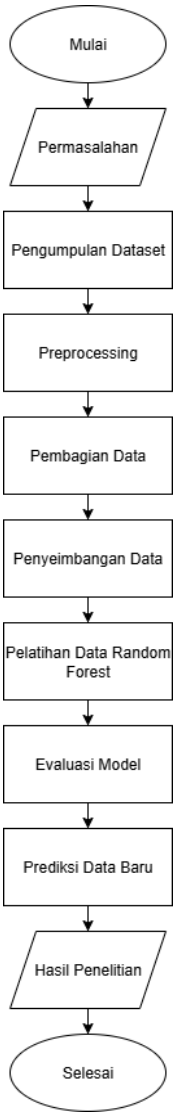
Machine Learning (ML) adalah pendekatan dari kecerdasan buatan yang memungkinkan sistem belajar dari data tanpa instruksi eksplisit. Dalam tugas klasifikasi, sering muncul tantangan ketidakseimbangan data yang dapat menyebabkan bias terhadap kelas mayoritas. Untuk mengatasi hal ini, digunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*), yang menghasilkan sampel sintetis untuk kelas minoritas agar distribusi lebih merata dan mengurangi risiko overfitting [10].

Secara umum, *Machine Learning* dibagi menjadi tiga jenis: *supervised learning, unsupervised learning,* dan *reinforcement learning*. Dalam *supervised learning*, algoritma seperti Random Forest, banyak digunakan karena keunggulan masing-masing dalam akurasi, efisiensi, dan kemampuan mengelola data kompleks [11]. *Random Fores*t menggabungkan prediksi dari banyak pohon keputusan untuk meningkatkan akurasi dan mengurangi overfitting, serta efektif menangani data berdimensi tinggi [12].

Berdasarkan latar belakang tersebut, pengembangan model prediksi risiko stres berbasis pola hidup menjadi penting. Dengan membandingkan hasil prediksi Machine Learning menggunakan algoritma *Random Forest* dan skor PSS, individu dapat memahami kebiasaan yang memicu stres dan mengetahui langkah preventif yang tepat. Selain itu, model ini dapat digunakan dalam aplikasi kesehatan digital yang memantau kondisi stres pengguna dan memberikan rekomendasi yang sesuai dengan kebutuhan masing-masing, guna menjaga kesehatan mental dan kualitas hidup.

II. METODE PENELITIAN

Pada Penelitian ini menggunakan pendekatan kuantitatif dengan algoritma Machine Learning untuk analisis data, yang efektif dalam klasifikasi dan prediksi dengan akurasi tinggi. Desain *cross-sectional* digunakan untuk menganalisis hubungan antara pola hidup dan tingkat stres pada satu titik waktu. Tahapan penelitian dapat dilihat pada Gbr 1.

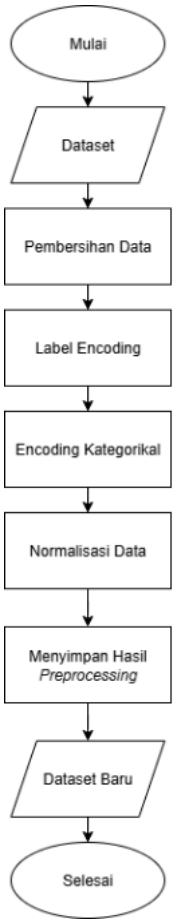


Gbr 1 Alur Penelitian

A. Pengumpulan Dataset Pola Hidup dan PSS

Data diperoleh melalui survei menggunakan *Google Form* dengan teknik *purposive sampling*. Responden berusia 13 - 45 tahun dari berbagai daerah di Indonesia dipilih berdasarkan variasi pola hidup dan tingkat stres yang relevan. Hanya data lengkap dan konsisten yang dianalisis. Kuesioner mencakup aspek pola hidup seperti durasi tidur, aktivitas fisik, konsumsi kafein, kebiasaan merokok, dan konsumsi alkohol. Instrumen yang digunakan diadaptasi dari alat ukur terstandar, yaitu *IPAQ-SF*, *PSQI*, pedoman Mayo Clinic, *AUDIT*, dan *Fagerstrom Test*. Tingkat stres diukur menggunakan *Perceived Stress Scale (PSS)-10*, yang terdiri dari 10 item dengan skala 0–4. Skor total diperoleh setelah membalik nilai item positif. Skor ≥20 menunjukkan stres berat. Skor dari *Perceived Stress Scale* (PSS) ini akan didapatkan melalui tes *Perceived Stress Scale* (PSS) secara online.

B. Preprocessing Data



Gbr 2 Alur Preprocessing

Setelah data terkumpul melalui survei melalui *Google Form*, file dengan format .csv tersebut menjalani tahapan *Preprocessing* untuk memastikan kualitas data yang dapat dilihat pada Gbr 2. Langkah-langkah yang dilakukan meliputi:

1. Pembersihan Data

   Data yang tidak lengkap dan duplikat dihapus. Nilai kosong ditangani dengan metode imputasi menggunakan rata-rata atau median, setelah terlebih dahulu diidentifikasi jumlahnya menggunakan fungsi .isnull().sum().

2. *Label Encoding*

   Skor *Perceived Stress Scale* (PSS) yang semula diklasifikasikan ke dalam tiga kategori (rendah, sedang, tinggi), diubah menjadi dua kelas: Stres (skor 27–40) dan Tidak Stres (skor 0–26). Penggabungan ini didasarkan pada ketidakseimbangan distribusi data, fokus penelitian terhadap stres berat, serta praktik umum dalam studi serupa. Transformasi ini juga memungkinkan model lebih fokus dalam mengidentifikasi faktor gaya hidup yang berkorelasi dengan stres tinggi.

3. Encoding Variabel Kategorikal

   Variabel kategorikal dari data gaya hidup dikonversi ke format numerik menggunakan *One-Hot Encoding*, agar dapat diproses oleh algoritma machine learning seperti Random Forest. Setiap kategori diubah menjadi representasi biner (0/1).

4. Normalisasi Data

   Seluruh variabel dinormalisasi menggunakan Min-Max Scaling agar berada dalam skala yang seragam. Ini penting untuk mencegah dominasi variabel tertentu karena perbedaan skala dalam proses pelatihan model.

C. Pembagian Data

Penelitian ini menerapkan metode *K-Fold Cross Validation* untuk menguji performa model secara menyeluruh dan mengurangi potensi bias akibat pembagian data yang tidak merata. Dibandingkan dengan pembagian data konvensional seperti train-test split, pendekatan ini memberikan estimasi performa yang lebih stabil dan generalisasi yang lebih baik. Dalam K-Fold Cross Validation, data dibagi menjadi k bagian berukuran sama. Model dilatih sebanyak k kali, setiap kali dengan satu fold sebagai data uji dan sisanya sebagai data latih. Hasil evaluasi dari tiap iterasi dirata-rata untuk memperoleh skor performa akhir. Penelitian ini membandingkan nilai k yang berbeda, yaitu 5, 10, dan 15, untuk melihat dampaknya terhadap akurasi dan kemampuan generalisasi model. Semakin besar nilai k, semakin banyak data yang digunakan untuk pelatihan, namun berdampak pada peningkatan waktu komputasi. Misalnya, pada 10-fold, model dilatih dengan 90% data dan diuji dengan 10% data sebanyak 10 kali. Dengan strategi ini, seluruh data digunakan bergantian sebagai data latih dan data uji, sehingga mengurangi risiko *overfitting* dan *underfitting*. Pendekatan ini memastikan bahwa model yang dibangun mampu bekerja secara konsisten terhadap data baru yang belum pernah dilihat sebelumnya.

D. Penyeimbangan Data

Untuk mengatasi permasalahan ketidakseimbangan kelas dalam variabel target, penelitian ini menerapkan metode *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE merupakan pendekatan oversampling yang menghasilkan data sintetis untuk kelas minoritas melalui interpolasi antar sampel yang ada, bukan dengan menduplikasi data secara langsung seperti pada metode konvensional. Teknik ini memungkinkan model untuk belajar dari representasi yang lebih bervariasi sehingga dapat mengurangi risiko

overfitting. Penerapan SMOTE dilakukan setelah proses encoding dan normalisasi, serta hanya diterapkan pada data latih (*training se*t) dalam setiap iterasi *cross-validation*. Hal ini bertujuan untuk menghindari *data leakage*, yakni masuknya informasi dari data uji ke dalam proses pelatihan yang dapat menyebabkan evaluasi performa model menjadi bias. Metode ini dipilih karena kemampuannya dalam menyeimbangkan distribusi kelas tanpa mengubah karakteristik kelas mayoritas [11]. Sejumlah studi sebelumnya juga menunjukkan bahwa SMOTE dapat meningkatkan performa model, terutama dalam metrik recall dan F1-score, pada kasus klasifikasi dengan distribusi kelas yang tidak seimbang secara signifikan [13].

E. Pelatihan Model

Pada tahap pelatihan, digunakan algoritma *Random Forest* untuk membangun model prediksi tingkat stres berdasarkan data pola hidup, yang mencakup durasi tidur, konsumsi kafein, aktivitas fisik, kebiasaan merokok, dan konsumsi alkohol. Label target diperoleh melalui instrumen *Perceived Stress Scale* (PSS), kemudian diklasifikasikan menjadi dua kategori, yaitu Stres dan Tidak Stres. Model dilatih untuk membedakan individu yang mengalami stres dari yang tidak, berdasarkan pola dalam data. *Random Forest* dipilih karena kemampuannya menangani data non-linear dan mengurangi overfitting melalui mekanisme ensambel, yaitu dengan membangun banyak pohon keputusan dan menggabungkan prediksinya melalui voting mayoritas. Selama pelatihan, dilakukan *tunning hyperparameter* guna mengoptimalkan performa model. Proses ini bertujuan untuk menemukan konfigurasi parameter yang memberikan akurasi dan generalisasi terbaik pada data uji.

F. Evaluasi Model

Tahap evaluasi dilakukan untuk mengukur performa model Random Forest dalam mengklasifikasikan tingkat stres individu ke dalam dua kategori, yaitu Stres dan Tidak Stres, berdasarkan pola hidup seperti durasi tidur, konsumsi kafein, aktivitas fisik, merokok, dan konsumsi alkohol. Evaluasi dilakukan menggunakan data uji yang belum pernah dilihat oleh model sebelumnya, guna mengukur kemampuan generalisasi. Evaluasi performa dilakukan dengan pendekatan *K-Fold Cross Validation*, sehingga seluruh data secara bergantian digunakan sebagai data latih dan data uji. Beberapa metrik utama yang digunakan dalam proses evaluasi adalah:

1. Akurasi: Mengukur proporsi prediksi yang benar terhadap seluruh prediksi.
2. Presisi: Menilai seberapa tepat model dalam memprediksi kelas "Stres".
3. Recall: Mengukur seberapa baik model mendeteksi semua kasus "Stres" yang benar.
4. F1-Score: Rata-rata harmonis antara presisi dan recall, memberikan evaluasi seimbang.

Nilai akhir dari masing-masing metrik dihitung dengan merata-ratakan hasil dari seluruh fold:

$$\bar{M} = \frac{1}{k}\sum_{i=1}^{k} M_i \quad (1)$$

Dengan menggunakan metrik-metrik tersebut, model dapat dievaluasi secara menyeluruh dan obyektif. Jika hasil evaluasi menunjukkan performa yang tinggi, maka dapat disimpulkan bahwa data pola hidup individu memiliki hubungan signifikan terhadap tingkat stres, serta model dapat diandalkan untuk tujuan prediktif atau intervensi dini.

G. Prediksi Data Baru

Berbeda dengan proses sebelumnya yang melibatkan perbandingan antara hasil prediksi dan label aktual, pada tahap ini data yang digunakan untuk prediksi tidak lagi menggunakan hasil dari PSS. Dengan demikian, model berperan sebagai alat inferensi, yang bertugas mengklasifikasikan tingkat stres individu berdasarkan pola hidupnya secara otomatis. Tahap ini merupakan penerapan nyata dari model machine learning dalam konteks dunia riil, di mana prediksi dilakukan terhadap data yang sepenuhnya baru dan tanpa label, dengan harapan bahwa hasilnya dapat memberikan nilai tambah dalam pengambilan keputusan atau strategi penanganan stres di masyarakat.

III. HASIL DAN PEMBAHASAN

Pada bagian ini menjelaskan hasil dari setiap tahapan implementasi yang telah dijelaskan pada Bab III sebelumnya. Proses pembangunan model prediksi tingkat stres menggunakan algoritma *Random Forest* diterapkan pada dataset yang telah dikumpulkan dan diproses oleh peneliti.

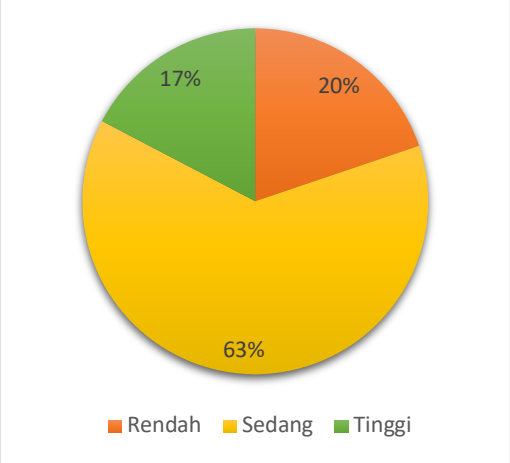
A. Pengumpulan Dataset Pola Hidup dan PSS

Pengumpulan data pada penelitian ini memanfaatkan data primer yang dikumpulkan melalui penyebaran kuesioner daring kepada responden berusia 13 hingga 45 tahun dari berbagai daerah. Kuesioner dalam penelitian ini dirancang untuk mengumpulkan data mengenai pola hidup responden serta hasil pengukuran tingkat stres berdasarkan tes PSS-10 yang dilakukan secara daring melalui website pengukuran

tingkat stres dengan PSS-10. Pendekatan yang digunakan bersifat kuantitatif, dengan tujuan untuk mengidentifikasi keterkaitan antara pola hidup dengan tingkat stres yang dialami individu. Data yang dikumpulkan selanjutnya digunakan untuk membangun model prediksi tingkat stres berbasis algoritma Machine Learning, khususnya dengan menggunakan model *Random Forest*. Data yang berhasil dikumpulkan sebanyak 204 data, yang mencakup variabel-variabel seperti frekuensi aktivitas berat dan sedang, durasi tidur, gangguan tidur, konsumsi dan reaksi terhadap kafein, serta kebiasaan merokok dan alkohol. Dari data yang diperoleh, dilakukan kategorisasi tingkat stres berdasarkan skor PSS-10 ke dalam tiga kelas, yaitu rendah, sedang, dan tinggi. Kategorisasi ini menjadi target atau label dalam proses klasifikasi menggunakan *Random Forest.* Distribusi tingkat stres berdasarkan hasil pengisian kuesioner oleh 204 responden dapat di lihat Tabel I dan Gbr 3.

Tabel I
distribusi responden

Tingkat Stres	Jumlah Responden	Persentase
Rendah	40	19.80%
Sedang	127	62.87%
Tinggi	35	17.33%
Total	202	100%



Gbr 3 hasil distribusi kuesioner

Distribusi yang relatif seimbang ini memungkinkan algoritma klasifikasi untuk belajar secara lebih efektif tanpa mengalami bias signifikan terhadap salah satu kelas. Namun, tetap dilakukan pemeriksaan distribusi kelas pada saat pelatihan model guna menghindari masalah data imbalance yang dapat mempengaruhi akurasi prediksi, terutama pada kelas minoritas.

B. Preprocessing Data

Tahapan berikutnya dalam penelitian ini adalah melakukan proses *preprocessing* terhadap data yang telah dikumpulkan melalui kuesioner. Data mentah yang diperoleh masih mengandung sejumlah permasalahan, seperti adanya nilai kosong *(missing values*) serta ketidakseimbangan pada distribusi kelas target. Oleh karena itu, dilakukan serangkaian langkah untuk meningkatkan kualitas data dan memastikan data tersebut siap digunakan dalam proses pelatihan model machine learning.

1. Pembersihan Data

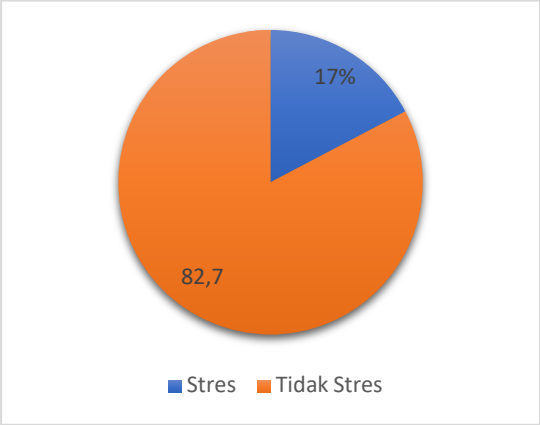
Fungsi *dropna()* digunakan untuk pada tahapan *preprocessing* ini untuk menghapus seluruh baris yang mengandung nilai kosong (*missing values*). Langkah ini diambil untuk menjaga integritas data serta menghindari potensi bias yang dapat muncul akibat proses imputasi, terutama pada data kategorikal ordinal. Dataset yang digunakan dalam penelitian ini terdiri dari 204 entri (responden) dan 11 atribut, yang mencakup 10 fitur input serta 1 variabel target berupa interpretasi tingkat stres. Ditemukan adanya dua entri yang mengandung nilai kosong di beberapa atribut. Keberadaan nilai kosong ini berpotensi mengganggu validitas analisis dan dapat memengaruhi kinerja model prediksi. Setelah dilakukan pembersihan, jumlah data yang siap digunakan dalam proses pelatihan model adalah sebanyak 202 entri.

2. *Label Encoding*

Pada langkah ini merubah data awal yang menunjukkan bahwa responden terbagi dalam tiga kategori tingkat stres berdasarkan hasil interpretasi PSS-10. Namun, merujuk pada pendekatan klasifikasi biner yang umum digunakan dalam penelitian sejenis, kategori tersebut kemudian digabung menjadi dua kelas utama, yaitu *Tidak Stres* dan *Stres*. Proses penggabungan dilakukan menggunakan fungsi *.replace()* dari pustaka *pandas.* Langkah penggabungan ini bertujuan untuk menyederhanakan proses klasifikasi sekaligus membantu mengurangi ketidakseimbangan distribusi antar kelas. Distribusi data setelah proses encoding label ditampilkan pada Tabel II dan Gbr 4.

Tabel II
distribusi responden 2 kelas

Kelas	Jumlah	Persentase
Tidak Stres	167	82.67%
Stres	35	17.33%
Total	202	100%



Gbr 4 distribusi responden 2 kelas

3. Encoding Variabel Kategorikal

Langkah ini dilakukan untuk mengubah variabel-variabel kategorikal pada data pola hidup menjadi format numerik yang dapat diproses oleh algoritma *machine learning*. Salah satu teknik yang digunakan adalah *one-hot encoding*, yaitu metode yang mengonversi setiap kategori unik dalam suatu variabel menjadi kolom biner terpisah. Dengan pendekatan ini, setiap nilai kategori direpresentasikan dalam bentuk angka 0 atau 1, sehingga dapat menghindari potensi bias yang mungkin timbul dari perbedaan skala atau urutan numerik yang tidak bermakna. Dari total 202 data yang telah melalui proses pembersihan, semula terdapat 11 kolom, terdiri dari 10 fitur input dan 1 target (label) yaitu interpretasi stres. Setelah dilakukan proses *one-hot encoding* terhadap fitur-fitur kategorikal, jumlah kolom meningkat secara signifikan menjadi 44 fitur. Peningkatan ini disebabkan oleh konversi setiap kategori unik dalam masing-masing fitur menjadi kolom tersendiri. Peningkatan jumlah fitur akibat proses one-hot encoding memang dapat memberikan representasi yang lebih akurat terhadap informasi kategorikal, namun juga berpotensi meningkatkan kompleksitas model dan risiko overfitting, terutama ketika jumlah data relatif kecil dibandingkan dengan jumlah fitur.

Oleh karena itu, penting untuk mempertimbangkan langkah-langkah lanjutan seperti seleksi fitur atau penerapan teknik regularisasi guna menjaga kinerja model tetap optimal. Selain itu, pemahaman terhadap makna tiap fitur yang dihasilkan juga diperlukan agar interpretasi hasil model tetap relevan dan sesuai konteks analisis stres berdasarkan pola hidup.

Dengan demikian, seluruh fitur kategorikal berhasil diubah ke dalam bentuk numerik, dan data menjadi siap digunakan dalam proses pelatihan model *Random Forest* seperti pada Gbr 5.

Contoh hasil fitur (X_balanced):

	aktivitas_berat_harian_1-3 hari	aktivitas_berat_harian_4-6 hari	aktivitas_berat_harian_7 hari	aktivitas_berat_harian_Tidak pernah	aktivitas_sedang_durasi_10-30 menit	aktivitas_sedang_durasi_31-60 menit	aktivitas_sedang_durasi_<10 menit	aktivitas_sedang_durasi_>60 menit
0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
1	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
2	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
3	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
4	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Gbr 5 hasil encoding kategorikal fitur

Untuk hasil encoding y seperti pada Gbr 6.

```
Mapping Label:
Stres => 0
Tidak Stres => 1

hasil y_encoded: [0 1 1 1 0 1 1 1 0 1]
```

Gbr 6 hasil encoding kategorikal label

4. Normalisasi Data

Setelah proses *encoding* selesai dilakukan, tahap berikutnya adalah normalisasi data menggunakan teknik *Min-Max Scaling*. Normalisasi ini bertujuan untuk memastikan bahwa seluruh fitur berada dalam rentang skala yang seragam, yaitu antara 0 dan 1. Hal ini penting agar algoritma *machine learning* tidak menjadi bias terhadap fitur yang memiliki skala numerik lebih besar dibandingkan fitur lainnya. Metode yang digunakan adalah *MinMaxScaler()* dari pustaka *scikit-learn*, yang bekerja dengan menghitung nilai minimum dan maksimum dari setiap kolom, kemudian mengubah setiap nilai fitur ke dalam rentang [0, 1]. Hasil dari proses ini menghasilkan data akhir sebanyak 202 baris dan 44 fitur numerik, di mana seluruh nilai sudah berada dalam skala seragam. Meskipun hasil dari *one-hot encoding* sudah menghasilkan nilai dalam rentang 0 dan 1, proses normalisasi tetap dilakukan untuk menjaga konsistensi

dalam pipeline *preprocessing*. Normalisasi ini juga berfungsi sebagai antisipasi apabila di kemudian hari terdapat fitur numerik tambahan (seperti usia, skor kebiasaan, atau nilai lainnya), sehingga keseluruhan data input memiliki skala yang seragam dan mendukung kinerja model yang optimal.

C. Pembagian Data

Setelah dilakukan preprocessing, dataset kemudian dibagi menggunakan metode *K-Fold Cross Validation*. Pada penelitian ini digunakan K-Fold dengan 5 fold, 10 fold, dan 15 fold, di mana data dibagi menjadi beberapa bagian yang masing-masing secara bergantian berperan sebagai data latih dan data uji. Proses ini dikonfigurasi dengan parameter *shuffle=True* untuk memastikan pembagian data secara acak, serta *random_state=42* agar hasilnya dapat direproduksi. Hasil pembagian data menunjukkan bahwa setiap fold memiliki proporsi data pelatihan dan pengujian yang relatif merata.

D. Penyeimbangan data

Setelah proses *preprocessing*, data dibagi menjadi data latih dan data uji menggunakan teknik *Stratified K-Fold Cross Validation* untuk menjaga proporsi distribusi kelas. Namun, hasil distribusi menunjukkan ketidakseimbangan kelas, baik pada klasifikasi dua label (Stres dan Tidak Stres). Ketidakseimbangan ini berisiko menyebabkan bias model terhadap kelas mayoritas. Oleh karena itu, digunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) untuk menyeimbangkan kelas hanya pada data latih, guna menghindari data leakage. SMOTE bekerja dengan menambahkan sampel sintetis pada kelas minoritas. hasil distribusi penyeimbangan data menggunakan SMOTE dilakukan dengan cara membuat sampel sintetis pada kelas minoritas, sehingga jumlahnya seimbang dengan kelas mayoritas seperti jika pada data training :

- Kelas "Stres" = 30 sampel
- Kelas "Tidak Stres" = 150 sampel

Maka SMOTE akan membuat 120 sampel buatan untuk kelas "Stres", sehingga keduanya jadi 150:150.
Dengan penerapan ini, distribusi menjadi seimbang dan model memiliki peluang yang lebih adil dalam mempelajari setiap kelas.

E. Hasil Implementasi Random Forest

Setelah dilakukan tahapan preprocessing seperti pembersihan data, *one-hot encoding,* normalisasi menggunakan *MinMaxScaler*, serta penyeimbangan data menggunakan metode SMOTE, algoritma *Random Forest* diterapkan untuk mengklasifikasikan tingkat stres berdasarkan pola hidup responden. Dalam proses pembangunan model prediksi tingkat stres, eksplorasi parameter atau *tunning hyperparameter* merupakan langkah penting untuk memperoleh performa model terbaik. Algoritma *Random Forest* memiliki sejumlah parameter yang dapat dikonfigurasi untuk meningkatkan akurasi dan generalisasi model terhadap data baru. Oleh karena itu, dilakukan eksplorasi parameter dengan pendekatan manual grid search menggunakan kombinasi parameter yang diuji satu per satu seperti pada Gbr 7.

```
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [5, 10, None],
    'min_samples_split': [2, 5],
    'class_weight': [None, 'balanced']
}
```

Gbr 7 tunning hyperparameter

Dari hasil eksplorasi parameter terdapat perbedaan dari setiap fold dengan 2 label yang ada dengan perbandingan hasil eksplorasi parameter seperti pada Tabel III.

Tabel III
hasil tunning hyperparameter random forest

Jumlah Fold	Parameter Terbaik
5-Fold	{'n_estimators': 100, 'max_depth': 5, 'min_samples_split': 2, 'class_weight': None}
10-Fold	{'n_estimators': 100, 'max_depth': 5, 'min_samples_split': 2, 'class_weight': None}
15-Fold	{'n_estimators': 200, 'max_depth': 5, 'min_samples_split': 2, 'class_weight': None}

Pemilihan parameter terbaik dilakukan berdasarkan evaluasi terhadap berbagai kombinasi parameter (lampiran). Parameter dipilih berdasarkan nilai F1-score tertinggi, karena metrik ini seimbang antara presisi dan recall, serta cocok untuk data tidak seimbang. Sebelum evaluasi, model dilatih menggunakan algoritma *Random Forest* yang dibangun dengan menginisialisasi beberapa parameter penting seperti n_estimators, max_depth, min_samples_split, dan class_weight. Parameter-parameter ini diperoleh dari hasil eksplorasi dan tuning untuk menghasilkan performa terbaik. Kemudian, model dilatih menggunakan data latih yang telah diseimbangkan dengan metode SMOTE. Proses pelatihan dilakukan dengan perintah *model.fit(X_train_sm, y_train_sm),* di mana model

mempelajari pola hubungan antara fitur-fitur gaya hidup dan tingkat stres responden. Langkah ini merupakan bagian penting dalam membentuk model prediksi yang andal sebelum dievaluasi pada data uji. Selanjutnya evaluasi model dengan menggunakan akurasi, presisi, recall, dan f-1 score dimana memiliki hasil yang berbeda dari 5-fold, 10-fold , dan 15-fold. Hasil evaluasi dari masing masing fold dengan dua label seperti pada tabel IV.

Tabel IV
perbandingan hasil 3 kfold

Jumlah Fold	Akurasi	Presisi	Recall	F-1 Score
5 Fold	0.81	0.86	0.92	0.89
10 Fold	0.80	0.86	0.90	0.88
15 Fold	0.83	0.87	0.93	0.90

Dari hasil yang diperoleh, terlihat bahwa 15-Fold menghasilkan akurasi tertinggi dibandingkan dengan 5-Fold dan 10-Fold. Hal ini dapat terjadi karena semakin besar jumlah fold, semakin banyak proporsi data yang digunakan untuk pelatihan pada setiap iterasi, sehingga model mendapatkan lebih banyak informasi dan dapat belajar pola dengan lebih baik. Selain itu, penggunaan lebih banyak fold juga memungkinkan evaluasi yang lebih stabil dan mendekati performa model terhadap data nyata. rata – rata akurasi menunjukkan performa yang konsisten dan relatif tinggi di setiap jumlah fold yang digunakan dalam proses *Stratified K-Fold Cross Validation.* Secara rinci, akurasi rata-rata pada 5-Fold sebesar 0.81, kemudian sedikit menurun menjadi 0.80 pada 10-Fold, dan kembali meningkat menjadi 0.83 pada 15-Fold. Hal ini menunjukkan bahwa model mampu melakukan generalisasi dengan baik meskipun jumlah fold bervariasi.

F. Prediksi Data Baru

Pengujian dilakukan untuk mengevaluasi sejauh mana model mampu memprediksi tingkat stres berdasarkan pola hidup responden. Model Random Forest terbaik hasil pelatihan sebelumnya disimpan dalam format .pkl menggunakan *joblib*, lalu dimuat kembali untuk keperluan prediksi. Data yang digunakan merupakan satu sampel dengan atribut serupa seperti data latih—meliputi durasi tidur, aktivitas fisik, konsumsi kafein, dan kebiasaan merokok. Data ini terlebih dahulu melalui proses pra-pemrosesan yang sama, yaitu encoding dan normalisasi dengan encoder serta scaler yang telah dibentuk selama pelatihan, guna memastikan kesesuaian struktur dan skala input dengan model. Prediksi dilakukan menggunakan fungsi *predict()* dari model *Random Forest*. Output yang dihasilkan berupa label stres atau tidak stres sesuai dengan pola hidup yang diberikan. Dengan pendekatan ini, model dapat digunakan langsung untuk klasifikasi tanpa memerlukan label selama data baru mengikuti format yang telah ditentukan. Uji coba dilakukan dengan memasukkan beberapa sampel data secara manual. Setiap sampel terdiri atas fitur-fitur yang sama seperti saat pelatihan. Model yang digunakan adalah hasil pelatihan dengan skema 15-fold cross-validation dengan dua label, yang sebelumnya menunjukkan performa terbaik. Oleh karena itu, hasil prediksi diharapkan mencerminkan kinerja optimal model.

1. Uji Coba Data Tidak Stres

Sampel pertama merupakan data pola hidup yang memenuhi kriteria untuk dilakukan prediksi tingkat stres menggunakan model Random Forest. Data tersebut ditampilkan pada Gbr 8.

```
data_baru = pd.DataFrame([{
    'aktivitas_berat_harian': '7 hari',
    'aktivitas_sedang_durasi': '31-60 menit',
    'durasi_tidur': '6-7 jam',
    'kesulitan_tidur': '1-2 kali seminggu',
    'konsumsi_kafein': '2-3 cangkir',
    'frekuensi_alkohol': 'Tidak pernah',
    'jumlah_alkohol': '<1 gelas',
    'jumlah_rokok': '≤10',
    'waktu_merokok_bangun_tidur': '>60 menit',
    'reaksi_kafein': 'Jarang'
}])
```

Gbr 8 data untuk hasil tidak stres

Dari data yang telah diinputkan memiliki output seperti pada Gbr 9.

Deskripsi	Hasil
Kategori Tingkat Stres	Tidak Stres

Gbr 9 hasil data tidak stres

2. Uji Coba Data Stres

Sampel kedua merupakan data pola hidup yang memenuhi kriteria untuk dilakukan prediksi tingkat stres menggunakan model Random Forest. Data tersebut ditampilkan pada Gbr 10.

```
data_baru = pd.DataFrame([{
    'aktivitas_berat_harian': 'Tidak pernah',
    'aktivitas_sedang_durasi': '<10 menit',
    'durasi_tidur': '<4 jam',
    'kesulitan_tidur': '5-7 kali seminggu',
    'konsumsi_kafein': '2-3 cangkir',
    'frekuensi_alkohol': '2-4 kali sebulan',
    'jumlah_alkohol': '4-5 gelas',
    'jumlah_rokok': '11-20',
    'waktu_merokok_bangun_tidur': '<6 menit',
    'reaksi_kafein': 'Sering'
}])
```

Gbr 10 data untuk hasil stres

Dari data yang telah diinputkan memiliki output seperti pada Gbr 11.

Deskripsi	Hasil
Kategori Tingkat Stres	Stres

Gbr 11 hasil data tidak stres

Hasil ini menunjukkan bahwa model Random Forest tidak hanya memiliki performa baik pada data pelatihan dan pengujian, tetapi juga memiliki kemampuan generalisasi yang baik terhadap data baru yang belum pernah dilihat sebelumnya.

IV. KESIMPULAN

Berdasarkan hasil penelitian mengenai prediksi tingkat stres menggunakan algoritma machine learning, khususnya Random Forest, dapat disimpulkan hal-hal berikut:

1. Skala *Perceived Stress Scale (PSS-10)* berhasil digunakan sebagai dasar dalam pelabelan tingkat stres responden. Data dikumpulkan melalui kuesioner dan diklasifikasikan ke dalam tiga kategori: rendah, sedang, dan tinggi. Untuk keperluan pemodelan, kategori ini disederhanakan menjadi dua kelas, stres dan tidak stres agar meningkatkan akurasi model. Hal ini menjadikan PSS-10 sebagai landasan yang valid dan terukur dalam membangun sistem prediksi tingkat stres.
2. Model *Random Forest* diterapkan pada data pola hidup responden, yang mencakup variabel seperti kualitas tidur, aktivitas fisik, konsumsi kafein,konsumsi alkohol dan kebiasaan merokok. Data diproses menggunakan *Stratified K-Fold Cross Validation* dan diseimbangkan dengan metode SMOTE. Proses *tuning hyperparameter* dilakukan untuk mengoptimalkan performa model.
3. Berdasarkan hasil evaluasi, *Random Forest* menunjukkan performa terbaik dengan klasifikasi dua label, dengan akurasi mencapai 0.83 dan F1-Score 0.90. Hasil ini menunjukkan bahwa algoritma ini mampu mengklasifikasikan tingkat stres secara efektif dan konsisten dalam konteks data yang digunakan.

Secara keseluruhan, hasil ini menunjukkan bahwa skala PSS dapat dikombinasikan dengan pendekatan machine learning, khususnya Random Forest, untuk membangun sistem prediksi tingkat stres berbasis pola hidup secara efektif. Model ini mampu mengidentifikasi kecenderungan stres berdasarkan kebiasaan hidup, dan dengan akurasi 83%, dapat dijadikan alat bantu awal dalam deteksi stres. Meski begitu, model ini belum dapat digunakan sebagai alat diagnosis tunggal tanpa validasi medis lebih lanjut.

DAFTAR PUSTAKA

[1] WHO Team, *Mental Health Atlas 2020*. 2021.

[2] Kemenkes RI, "Kemenkes Beberkan Masalah Permasalahan Kesehatan Jiwa di Indonesia – Sehat Negeriku," *Kemenkes RI*, pp. 7–10, 2021, [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20211007/1338675/kemenkes-beberkan-masalah-permasalahan-kesehatan-jiwa-di-indonesia/.

[3] R. Setiawan and S. Halim, "Hubungan aktivitas fisik dan tingkat stres pada mahasiswa fakultas kedokteran universitas tarumanagara," *JKKT J. Kesehat. dan Kedokt. Tarumanagara*, vol. 2, no. 1, pp. 16–19, 2023.

[4] P. Columbia, "Dampak Kurang Tidur Dalam Psikologi," pp. 1–7, [Online]. Available: https://www.perplexity.ai/search/dampak-kurang-tidur-t.7B9fl_Roumf0vg0zlNLw.

[5] E. Scott, "Caffeine, Stress and Your Health," *VerywellMind*, pp. 1–4, 2020, [Online]. Available: https://www.verywellmind.com/caffeine-stress-and-your-health-3145078.

[6] D. Schultchen *et al.*, "Bidirectional relationship of stress and affect with physical activity and healthy eating," *Br. J. Health Psychol.*, vol. 24, no. 2, pp. 315–333, 2019, doi: 10.1111/bjhp.12355.

[7] Z. Zakiyah, Y. A. Sihombing, M. I. Kamaruddin, G. A. Salomon, and M. Anshari, "Stress Level and Smoking Behavior," *J. Ilm. Kesehat. Sandi Husada*, vol. 12, no. 2, pp. 467–473, 2023, doi: 10.35816/jiskh.v12i2.1118.

[8] M. W. Manoppo, F. F. Pitoy, and K. B. Tampi, "Hubungan Tingkat Stres dengan Konsumsi Alkohol pada Remaja," *MAHESA Malahayati Heal. Student J.*, vol. 3, no. 6, pp. 1710–1725, 2023, doi: 10.33024/mahesa.v3i6.10585.

[9] K. M. Harris, A. E. Gaffey, J. E. Schwartz, D. S. Krantz, and M. M. Burg, "The Perceived Stress Scale as a Measure of Stress: Decomposing Score Variance in Longitudinal Behavioral Medicine Studies," *Ann. Behav. Med.*, vol. 57, no. 10, pp. 846–854, 2023, doi: 10.1093/abm/kaad015.

[10] E. A. Elsoud *et al.*, "Under Sampling Techniques for Handling Unbalanced Data with Various Imbalance Rates: A Comparative Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, pp. 1274–1284, 2024, doi: 10.14569/IJACSA.2024.01508124.

[11] Wijoyo A, Saputra A, Ristanti S, Sya'ban S, Amalia M, and Febriansyah R, "Pembelajaran Machine Learning," *OKTAL (Jurnal Ilmu Komput. dan Sci.*, vol. 3, no. 2, pp. 375–380, 2024, [Online]. Available: https://journal.mediapublikasi.id/index.php/oktal/article/view/2305.

[12] 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.

[13] Y. Liu, C. M. Eckert, and C. Earl, "A review of fuzzy AHP methods for decision-making with subjective judgements," *Expert Syst. Appl.*, vol. 161, no. December, pp. 1–10, 2020, doi: 10.1016/j.eswa.2020.113738.