BAB I

PENDAHULUAN

1.1 Latar Belakang

Kesehatan mental, khususnya depresi, telah menjadi salah satu tantangan global paling krusial yang memengaruhi produktivitas dan kualitas hidup individu di berbagai kelompok demografis [1]. Fenomena ini tidak lagi dipandang sekadar sebagai gangguan psikologis murni, melainkan sebuah manifestasi kompleks yang dipicu oleh interaksi berbagai faktor multidimensional [2]. Di era modern, determinan utama yang berkontribusi signifikan terhadap peningkatan risiko depresi meliputi tingginya tekanan kerja (*work pressure*) dan tekanan akademik (*academic pressure*), yang sering kali berjalan beriringan dengan penurunan durasi tidur serta memburuknya kebiasaan diet atau pola makan. Secara biologis, akumulasi stres kronis akibat beban profesional dan akademik ini terbukti mampu memicu disregulasi sistem saraf pusat, mengganggu plastisitas sinaptik, bahkan mempercepat penuaan biologis pada tingkat seluler yang bermuara pada depresi berat [3]. Oleh karena itu, kemampuan untuk memprediksi dan memetakan risiko depresi berdasarkan variabel-variabel gaya hidup dan tekanan psikososial tersebut menjadi sangat urgensi untuk dilakukan.

Faktor pertama yang memiliki kontribusi signifikan terhadap kerentanan psikologis adalah tekanan kerja (*work pressure*). Di era profesional modern yang menuntut produktivitas tinggi, beban kerja yang berlebihan, tenggat waktu yang ketat, serta ekspektasi performa yang tidak realistis sering kali memicu stres kerja kronis. Akumulasi dari tekanan profesional ini secara bertahap menguras sumber daya kognitif dan emosional individu. Jika kondisi ini terjadi berkepanjangan tanpa adanya mekanisme koping (*coping mechanism*) yang efektif, stres tersebut akan berkembang menjadi kejenuhan kerja (*burnout*) tingkat tinggi yang secara klinis terbukti mampu memicu timbulnya gejala depresi berat [4]. Di sisi lain, ketidakseimbangan antara usaha yang dikeluarkan dengan penghargaan yang diterima di lingkungan kerja juga terbukti

memperparah penurunan motivasi kerja dan memicu alienasi sosial yang berujung pada distres emosional mendalam [5].

Selain lingkungan profesional, tekanan akademik (*academic pressure*) juga menjadi determinan utama, terutama pada kelompok individu yang berada dalam masa pendidikan. Tuntutan untuk mempertahankan standar nilai yang tinggi, kompetisi ketat antar rekan sejawat, beban tugas yang masif, hingga kecemasan dalam menghadapi ujian atau menyelesaikan tugas akhir menciptakan beban mental yang konstan . Tekanan psikologis di lingkungan akademis ini sering kali memicu kecemasan akut dan rasa tidak berdaya yang mengganggu stabilitas emosional, sehingga memperbesar peluang terjadinya depresi secara signifikan akibat kelelahan mental yang berkepanjangan [6]. Tekanan konstan ini jika tidak dimitigasi juga berkorelasi erat dengan penurunan efikasi diri akademik, yang lambat laun memicu keputusasaan klinis dan penarikan diri dari lingkungan social [7].

Faktor demografis seperti usia (*age*) juga memegang peranan krusial dalam memetakan tingkat kerentanan terhadap depresi. Setiap tahapan usia membawa beban transisi, tanggung jawab sosial, dan tantangan psikososial yang berbeda, seperti masa dewasa awal yang dipenuhi ketidakpastian masa depan, pencarian jati diri, hingga adaptasi dunia kerja. Karakteristik biologis, maturasi sistem saraf, serta kematangan emosional yang bervariasi di setiap rentang usia menyebabkan respons individu terhadap stresor lingkungan menjadi tidak seragam, sehingga menjadikan faktor usia sebagai variabel prediktor penting dalam mengidentifikasi kelompok populasi yang paling rentan [8]. Pergeseran neurobiologis seiring bertambahnya usia, dikombinasikan dengan akumulasi stresor lingkungan hidup, secara signifikan mengubah ambang batas toleransi individu terhadap kecemasan kronis [9].

Ditinjau dari aspek fisiologis dan pola hidup, kebiasaan diet (*dietary habits*) atau pola makan memiliki hubungan dua arah yang sangat erat dengan kesehatan mental. Konsumsi makanan yang tidak seimbang, kekurangan nutrisi mikro esensial, serta pola makan yang tidak teratur akibat stres dapat

mengganggu fungsi poros usus-otak (*gut-brain axis*). Disregulasi pada sistem pencernaan dan mikrobioma usus ini terbukti memengaruhi produksi serta transmisi neurotransmiter penting seperti serotonin dan dopamin yang bertanggung jawab atas regulasi suasana hati (*mood*), sehingga memperburuk kondisi psikologis dan meningkatkan kerentanan terhadap depresi [10]. Selain itu, pola konsumsi tinggi lemak jenuh dan gula rafinasi terbukti memicu inflamasi sistemik tingkat rendah yang dapat menembus sawar darah otak, mengganggu fungsi kognitif, dan merusak stabilitas emosi [11].

Variabel gaya hidup lain yang memiliki dampak langsung terhadap stabilitas mental adalah durasi tidur (*sleep duration*). Tidur merupakan proses biologis yang sangat krusial untuk restorasi fungsi kognitif, konsolidasi memori, serta regulasi emosi harian, namun sering kali dikorbankan demi memenuhi tuntutan aktivitas yang padat. Kurangnya waktu tidur secara kronis menyebabkan gangguan pada ritme sirkadian tubuh dan memicu lonjakan hormon kortisol atau hormon stres. Kondisi deprivasi tidur ini tidak hanya menurunkan kemampuan konsentrasi dan pengambilan keputusan, tetapi juga secara langsung memperlemah ketahanan psikologis individu dalam menghadapi tekanan emosional sehari-hari [12]. Gangguan tidur yang berkepanjangan ini pada akhirnya merusak konektivitas fungsional antara amigdala dan korteks prefrontal, yang secara klinis bermanifestasi sebagai ketidakstabilan suasana hati yang ekstrem dan peningkatan skor keparahan depresi [13].

Dalam upaya melakukan deteksi dini dan klasifikasi risiko depresi berdasarkan kelima variabel tersebut, pendekatan berbasis *machine learning* menawarkan efisiensi dan akurasi yang jauh lebih tinggi dibandingkan dengan metode diagnostik konvensional [14]. Untuk mengatasi kompleksitas hubungan non-linear antar variabel prediktor tersebut, algoritma Random Forest dipilih sebagai model dasar (*baseline*) dalam penelitian ini. Berdasarkan peta perkembangan teknologi (*state of the art*) dalam klasifikasi data tabular empiris, Random Forest secara konsisten menunjukkan performa superior jika dibandingkan dengan model tradisional seperti *Logistic Regression* atau *Single*

Decision Tree [15]. Keunggulan utama Random Forest terletak pada pendekatan *ensemble learning* berbasis *bagging*, yang membangun puluhan hingga ratusan pohon keputusan secara independen melalui teknik substitusi acak [16]. Karakteristik ini membuat Random Forest sangat tangguh (*robust*) terhadap *outliers* dan noise data, serta meminimalkan risiko *overfitting* yang sering terjadi pada dataset multidimensi [17]. Selain itu, algoritma ini memiliki kemampuan inheren untuk mengukur tingkat kepentingan fitur (*feature importance*), suatu aspek yang sangat krusial untuk mengevaluasi variabel mana yang paling dominan dalam memicu depresi [18].

Namun, kendala utama yang sering dihadapi dalam pemodelan data kesehatan mental yang bersumber dari survei atau observasi lapangan adalah masalah ketidakseimbangan kelas (*class imbalance*) [19]. Jumlah individu yang didiagnosis mengalami depresi klinis (kelas minoritas) umumnya jauh lebih sedikit dibandingkan dengan populasi yang berada dalam kondisi normal atau sehat (kelas mayoritas). Jika data yang timpang ini langsung dimasukkan ke dalam algoritma klasifikasi tanpa penanganan khusus, model akan mengalami bias yang tinggi terhadap kelas mayoritas [20]. Model cenderung menghasilkan akurasi semu yang tinggi secara keseluruhan, tetapi gagal total (*poor sensitivity*) dalam mendeteksi individu yang sebenarnya benar-benar mengalami depresi. Meskipun Random Forest memiliki arsitektur yang kuat, algoritma ini tetap rentan terhadap penurunan performa jika dihadapkan pada *class imbalance*ekstrem, karena kriteria pemisahan (*splitting criteria*) seperti *Gini Impurity* secara inheren didorong oleh minimalisasi kesalahan klasifikasi global [21]. Oleh karena itu, diperlukan strategi optimasi pada tingkat data (*data-level approach*) sebelum proses pelatihan model dilakukan, dengan menetapkan Synthetic Minority Over-sampling Technique (SMOTE) sebagai solusi optimasi tunggal karena keunggulannya yang jauh lebih stabil dibandingkan metode lainnya.

Sebagai perbandingan, metode alternatif seperti *Random Over-Sampling* (ROS) bekerja dengan cara mereplikasi atau menggandakan sampel kelas minoritas secara acak. Pendekatan ROS ini menyebabkan ruang keputusan

(*decision boundary*) yang dibentuk oleh Random Forest menjadi terlalu sempit dan kaku, yang pada akhirnya memicu terjadinya *overfitting* yang parah pada data uji karena model hanya menghafal data tiruan yang persis sama [22]. Sebaliknya, SMOTE tidak melakukan duplikasi, melainkan menciptakan sampel sintetis baru di sepanjang garis acak yang menghubungkan beberapa *k-nearest neighbors* dari kelas minoritas, sehingga mampu memperluas ruang keputusan model secara lebih general dan realistis [23]

Di sisi lain, penggunaan *Random Under-Sampling* (RUS) juga dinilai kurang ideal karena menyelesaikan masalah ketidakseimbangan dengan menghapus sebagian besar sampel dari kelas mayoritas secara acak agar setara dengan kelas minoritas. Strategi RUS ini sangat merugikan dalam pemodelan data kesehatan mental karena menyebabkan hilangnya informasi berharga (*loss of valuable information*) terkait variasi karakteristik populasi sehat, yang justru sangat dibutuhkan oleh Random Forest untuk mengenali batas pemisah antar kelas secara utuh [24]. Sementara itu, algoritma *Adaptive Synthetic Sampling*(ADASYN) yang merupakan pengembangan dari SMOTE juga memiliki kelemahan mendasar berupa sensitivitas yang terlalu tinggi terhadap *outliers* atau data *noise* yang berada di sekitar perbatasan kelas. Hal ini sering kali menyebabkan ADASYN menciptakan sampel sintetis di lokasi yang keliru, yang mengaburkan ruang keputusan Random Forest dan menurunkan metrik nilai *Precision* secara signifikan [25] [depression 30]. Dengan mengintegrasikan SMOTE, distribusi kelas dapat disetarakan secara proporsional tanpa risiko kehilangan informasi esensial dan tanpa menjebak model dalam kondisi *overfitting*.

Berdasarkan urgensi dan dinamika teknis tersebut, penelitian ini dilakukan untuk mengevaluasi, menganalisis, dan membandingkan performa klasifikasi antara algoritma Random Forest konvensional dengan Random Forest yang telah dioptimasi menggunakan SMOTE. Melalui perbandingan komparatif ini, diharapkan dapat ditemukan model prediktif terbaik yang tidak hanya unggul pada akurasi global, melainkan juga memiliki sensitivitas dan nilai *F1-Score* yang optimal dalam memetakan risiko depresi berdasarkan variabel tekanan

kerja, tekanan akademik, usia, kebiasaan diet, serta durasi tidur [26]. Validasi model yang komprehensif ini sangat krusial untuk menghasilkan sistem skrining awal yang andal, sehingga dapat mendukung pengambilan keputusan klinis atau intervensi preventif yang lebih tepat sasaran di masyarakat [27].

1.2 Rumusan Masalah

1. Bagaimana tingkat bias performa dan keterbatasan algoritma *Random Forest* murni (*baseline*) ketika dihadapkan pada dataset risiko depresi yang memiliki ketidakseimbangan proporsi data sebesar 82% kelas sehat dan 18% kelas depresi?
2. Bagaimana pengaruh penerapan teknik *oversampling Synthetic Minority Over-sampling Technique* (SMOTE) terhadap peningkatan metrik sensitivitas (*Recall*) pada algoritma *Random Forest* dalam mendeteksi kelas minoritas risiko depresi?
3. Bagaimana perbandingan performa antara model *Random Forest* murni dan *Random Forest* + SMOTE berdasarkan metrik evaluasi *Accuracy*, *Precision*, dan *Recall* demi menghasilkan model klasifikasi yang paling efisien dan aman untuk penapisan kesehatan mental?

1.3 Batasan Masalah

1. Penelitian ini menggunakan dataset survei sekunder terkait kesehatan mental dan gaya hidup yang telah tersedia, sehingga peneliti tidak melakukan pengumpulan data primer atau wawancara baru secara langsung.
2. Variabel independen yang dianalisis dibatasi pada tiga aspek gaya hidup dan ekonomi, yaitu: tekanan akademik, tekanan pekerjaan, umur, durasi tidur, pola makan, jam kerja/belajar (*work/study hours*), dan tingkat stres finansial.

3. Variabel dependen adalah kondisi depresi yang didefinisikan dalam bentuk klasifikasi biner, di mana nilai 0 merepresentasikan kondisi tidak depresi dan nilai 1 merepresentasikan kondisi depresi.
4. Metode pemodelan yang digunakan dibatasi pada dua algoritma utama:

    a. Random Forest Standar yang diposisikan sebagai model dasar (*baseline model*) tanpa penanganan ketidakseimbangan data.
    b. Random Forest (RF) yang diintegrasikan dengan teknik SMOTE (Synthetic Minority Over-sampling Technique) sebagai model yang dioptimasi untuk menangani ketidakseimbangan data (class imbalance).

5. Evaluasi performa model hanya dilakukan berdasarkan metrik standar machine learning, yaitu: Accuracy, Precision, Recall, F1-score, dan Area Under the ROC Curve (AUC-ROC).
6. Penelitian ini terbatas pada analisis komparatif performa algoritma secara teknis dan tidak bertujuan untuk memberikan diagnosis medis klinis terhadap individu secara langsung.
7. Penelitian ini tidak mencakup tahap *deployment* atau implementasi sistem secara nyata (aplikasi/produk siap pakai). Tahap *Deployment* dalam kerangka kerja CRISP-DM pada penelitian ini hanya berupa rekomendasi konseptual bagi pengembangan sistem skrining di masa mendatang, bukan bagian dari hasil penelitian yang diimplementasikan.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

1. Menganalisis kelemahan serta dampak dari ketidakseimbangan data terhadap bias prediksi algoritma *Random Forest* murni sebelum dilakukan intervensi data.

2. Menerapkan teknik SMOTE pada fase pra-pemrosesan data untuk menyeimbangkan proporsi kelas minoritas sehingga meningkatkan kepekaan algoritma terhadap karakteristik data risiko depresi.
3. Mengevaluasi dan membandingkan secara empiris performa model *Random Forest* murni (*baseline*) dengan *Random Forest* + SMOTE (*proposed model*) untuk membuktikan bahwa intervensi penyeimbangan data mampu mengoptimalkan nilai *Recall* demi menekan angka kegagalan deteksi risiko depresi.

1.4.2 Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis, sebagai berikut:

a. Memberikan kontribusi dalam pengembangan literatur terkait penerapan machine learning di bidang psikologi kesehatan, khususnya dalam penggunaan algoritma ensemble untuk klasifikasi depresi.
b. Optimalisasi Data: Memberikan kajin ilmiah mengenai efektivitas teknik Synthetic Minority Over-sampling Technique (SMOTE) dalam menangani masalah ketidakseimbangan data (class imbalance) pada dataset kesehatan mental.
c. Referensi Akademik: Menjadi referensi atau landasan bagi peneliti selanjutnya yang ingin mengembangkan model prediksi depresi menggunakan variabel gaya hidup dan faktor ekonomi.
d. Memberikan masukan dalam pengembangan alat skrining kesehatan mental yang praktis, murah, dan mudah diakses oleh masyarakat luas karena hanya berbasis pada variabel perilaku harian.

e. Pengambilan Keputusan: Membantu praktisi kesehatan atau institusi pendidikan dalam mengidentifikasi individu yang berisiko tinggi mengalami depresi secara lebih akurat, sehingga intervensi dapat dilakukan lebih cepat.
f. Pemetaan Faktor Risiko: Memberikan gambaran mengenai faktor mana yang paling dominan (seperti durasi tidur atau stres finansial) dalam memengaruhi risiko depresi, sehingga edukasi kesehatan dapat dilakukan secara lebih tertarget.

1.5 Sistematika Penulisan

BAB I PENDAHULUAN Bab pertama pada penelitian ini berisikan latar belakang yang menguraikan urgensi permasalahan depresi dan pengaruh variabel gaya hidup serta tekanan lingkungan, yang kemudian membentuk rumusan masalah, tujuan penelitian, batasan penelitian, serta manfaat penelitian.

BAB II LANDASAN TEORI Pada bab kedua berisikan kajian literatur yang menjelaskan konsep, teori, dan elemen terkait penelitian, termasuk penjelasan mengenai gangguan depresi, variabel penelitian (*work pressure*, *academic pressure*, *age*, *dietary habits*, *sleep duration*, dan *financial stress*), serta teori algoritma *Machine Learning* yang digunakan.

BAB III METODOLOGI PENELITIAN Pada bab ketiga, dijelaskan secara detail mengenai metodologi penelitian yang digunakan, prosedur pengumpulan data, penjelasan terkait fitur-fitur dalam dataset, serta tahapan alur penelitian mulai dari persiapan hingga pengujian model.

BAB IV ANALISIS DAN HASIL PENELITIAN Pada bab keempat berisikan detail prosedur penelitian menggunakan kerangka kerja CRISP-DM dalam melakukan klasifikasi risiko depresi. Bab ini mencakup tahapan *Data Understanding*, *Data Preparation* (termasuk penerapan SMOTE), hingga proses pemodelan menggunakan *Random Forest*. Di akhir bab, ditampilkan

hasil evaluasi akurasi model yang dibandingkan dengan hasil penelitian terkait serta diskusi mendalam mengenai faktor dominan pemicu depresi.

BAB V SIMPULAN DAN SARAN Pada bab kelima menampilkan hasil akhir yang menjawab tujuan penelitian secara ringkas dan spesifik berdasarkan temuan pada bab sebelumnya. Bab ini juga menyertakan saran konstruktif untuk pengembangan instrumen deteksi dini maupun perbaikan metode pada penelitian kesehatan mental di masa mendatang.

BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Tabel 2. 1 Penelitian Terdahulu

Jurnal	Masalah	Metode	Hasil	Limitasi
[19]	Kesulitan dalam mendeteksi risiko depresi pada kelompok minoritas akibat adanya ketidakseimbangan data (*imbalanced data*) dalam survei kesehatan mental skala besar.	*Random Forest* + SMOTE	Penggunaan teknik penyeimbangan data (SMOTE) terbukti meningkatkan akurasi model secara drastis dari 72% menjadi 89,2%.	Model ini lebih terfokus pada perilaku pencarian di internet, sehingga kurang melibatkan variabel lingkungan kerja atau gaya hidup fisik.
[9]	Kebutuhan untuk mengukur tingkat keparahan depresi menggunakan kuesioner	Algoritma *Machine Learning* (*Random Forest*)	Model *Random Forest* berhasil mengklasifikasikan tingkat risiko depresi dengan akurasi mencapai 85,4%.	Data yang digunakan kurang memiliki penanganan khusus untuk menyeimbangkan kelas

	multidimensi yang memiliki variabel beraneka ragam.			data minoritas (responden depresi).
[1]	Memahami dampak profil risiko sosial (seperti beban ekonomi dan tekanan kerja) terhadap penuaan biologis dan depresi pada populasi nasional berskala besar.	*Logistic Regression* pada *Big Data* (NHANES)	Model regresi memiliki performa diskriminasi yang kuat dalam membedakan kelas depresi dengan nilai AUROC 0,78(setara performa ~78%).	Menggunakan desain *cross-sectional* sehingga tidak dapat membuktikan hubungan sebab-akibat secara absolut.
[27]	Pendekatan model "satu untuk semua" kurang akurat dalam memprediksi depresi pada kelompok spesifik yang	*Hierarchical Bayesian* & *Machine Learning*	Model mencapai tingkat akurasi 80,5% dengan kemampuan prediksi yang sangat sensitif pada subgrup usia mahasiswa.	Hasil pemodelan terlalu spesifik pada lingkungan institusi pendidikan, sehingga generalisasi

	rentan tekanan akademik tinggi.			pada populasi pekerja umum terbatas.
[20]	Upaya mengidentifikasi indikator kecemasan dan depresi tersembunyi melalui pola komunikasi digital dan beban mental pengguna.	Analisis Sentimen & *Ensemble Learning* (*Random Forest*)	Mampu mendeteksi tanda-tanda depresi berdasarkan rekam jejak linguistik pengguna dengan tingkat akurasi sebesar 82,1%.	Pemodelan sangat bergantung pada tingkat keaktifan individu di media sosial, mengabaikan mereka yang pasif secara digital.
[25]	Evaluasi dampak pergeseran gaya hidup seperti kualitas durasi tidur dan kebiasaan sarapan terhadap peningkatan tren depresi selama 10	*Complex Samples Multiple Logistic Regression*	Algoritma tervalidasi sangat baik (*Odds Ratio* prediktif tinggi) pada analisis *big data* yang mencakup 536.450 data responden.	Mengandalkan data kuesioner pelaporan mandiri (*self-reported*) yang rentan terhadap bias ingatan responden jangka panjang.

	tahun terakhir.			
[28]	Keterbatasan prosedur medis konvensional untuk memonitor korelasi antara stabilitas emosional dengan gangguan pola tidur.	Algoritma Prediktif pada *Microblog Big Data*	Berhasil memprediksi gangguan tidur sebagai indikator depresi dengan akurasi prediktif sekitar 83% berdasarkan jejak bahasa.	Ekspresi emosi di ruang publik digital mungkin tidak sepenuhnya merefleksikan keparahan kondisi klinis yang sebenarnya.
[24]	Menganalisis pengaruh lingkungan kerja (*occupational exposure*) yang buruk terhadap depresi pekerja pada populasi skala nasional.	*Logistic Regression* pada *Big Data* (55.649 pekerja)	Model regresi berhasil mengidentifikasi dan memprediksi korelasi beban kerja dengan gangguan mental dengan signifikansi yang tinggi.	Variabel pemodelan berpusat pada bahaya fisik pekerjaan dan kurang mengeksplorasi variabel internal seperti asupan gizi (*dietary*).
[14]	Terbatasnya akses layanan pakar untuk	Sistem Pakar berbasis *Mac*	Model mengintegrasikan multivariabel input	Model cenderung kaku dan

	mendiagnosis depresi ketika pasien memiliki kombinasi gejala dan variabel pemicu yang rumit.	*hine Learning*	dan menghasilkan diagnosis klasifikasi dengan akurasi mencapai 87%.	membutuhkan pembaruan parameter manual dari pakar saat ada tren gaya hidup baru.
[29]	Identifikasi dini perilaku depresif dari jejak interaksi pada komunitas online sebelum berkembang menjadi depresi klinis berat.	*Data Mining* & Algoritma Klasifikasi	Algoritma mampu mengekstraksi fitur perilaku dan mengidentifikasi penderita depresi dengan tingkat akurasi sekitar 81%.	Kesulitan dalam memverifikasi secara pasti kebenaran data demografis (seperti usia dan profesi) dari pengguna yang anonim.

Penelitian ini dibangun berdasarkan evolusi metodologi deteksi kesehatan mental, yang terus bergeser dari pendekatan statistik konvensional menuju pemodelan prediktif berbasis *machine learning*. Sejumlah studi terdahulu sering kali mengandalkan metode linier, seperti regresi logistik, untuk memetakan determinan sosial dan beban kerja pada populasi berskala besar [30], [31]. Meskipun pendekatan awal tersebut mampu memberikan gambaran peluang (*odds ratio*) dari variabel prediktor seperti tingkat stres finansial atau

durasi tidur [18], metode linier memiliki keterbatasan fundamental dalam menangkap pola interaksi non-linier yang kompleks pada data perilaku manusia. Hal ini mendorong transisi penggunaan algoritma *ensemble* seperti *Random Forest* sebagai standar dasar (*baseline*), yang terbukti secara empiris lebih tangguh dan stabil dalam menangani variabilitas data gaya hidup multidimensi tanpa terjebak pada asumsi linieritas [12], [19].

Perbandingan performa antar literatur menunjukkan bahwa algoritma klasifikasi yang tangguh sekalipun tetap rentan mengalami bias ketika dihadapkan pada dataset kesehatan mental yang sangat tidak seimbang (*imbalanced*). Untuk mengatasi dominasi kelas mayoritas (responden sehat), penerapan teknik optimasi pada tingkat data seperti SMOTE (*Synthetic Minority Over-sampling Technique*) muncul sebagai solusi yang paling komprehensif. Efektivitas ini dibuktikan melalui peningkatan performa model secara drastis, di mana integrasi SMOTE pada algoritma *Random Forest*mampu mendongkrak akurasi klasifikasi dari 72% menjadi 89,2% [32]. Keunggulan kombinasi *Random Forest* dan SMOTE ini terbukti mampu melampaui performa pemodelan probabilitas tingkat lanjut maupun sistem pakar konvensional dalam mengenali pola-pola halus dari gejala depresi pada kelompok berisiko tinggi [15], [22]. Metodologi berbasis penyeimbangan data sintetis dan pembelajaran *ensemble* ini kini diakui sebagai arsitektur terbaik untuk meminimalkan tingkat kesalahan deteksi (*false negative*).

Penelitian ini secara strategis menghubungkan berbagai temuan empiris tersebut dengan menetapkan *Random Forest*murni sebagai model dasar (*baseline model*), yang kemudian dikomparasikan langsung dengan *Random Forest* yang dioptimasi menggunakan SMOTE. Berbeda dengan penelitian sebelumnya yang cenderung mengandalkan satu dimensi data, seperti analisis linguistik di media sosial [19], [24] atau mengandalkan algoritma linier untuk data survey [18], [30] penelitian ini menawarkan kebaruan melalui studi isolasi (*ablation study*). Pendekatan ini secara spesifik mengukur seberapa besar dampak intervensi SMOTE terhadap algoritma *Random Forest* dalam

memproses kombinasi variabel tekanan harian (*work & academic pressure*) dan gaya hidup fisik. Fokus utamanya adalah membuktikan bahwa sinergi penyeimbangan data pada algoritma berbasis pohon keputusan (*decision tree*) mampu memberikan sensitivitas deteksi yang jauh lebih tajam dan adil dibandingkan model *baseline* murni, guna menghasilkan sistem penapisan dini depresi yang lebih responsif terhadap dinamika kesehatan mental saat ini.

2.2 Teori yang berkaitan

Landasan teori dalam penelitian ini berfungsi sebagai kerangka konseptual untuk memetakan hubungan antara determinan gaya hidup, tekanan lingkungan, dan kondisi kesehatan mental individu. Penjelasan ini mencakup variable-variabel kunci yang digunakan sebagai fitur prediktor serta kerangka kerja metodologis yang diterapkan.

2.2.1 Konsep Depresi dan Faktor Risiko Terkait

Depresi merupakan gangguan kesehatan mental yang ditandai dengan gangguan suasana hati, kehilangan minat, serta penurunan energi yang signifikan[17]. Secara klinis, depresi dapat dipicu oleh interaksi kompleks antara faktor biologis dan perilaku harian.

a. Durasi Tidur: Kualitas dan kuantitas tidur memiliki hubungan dua arah dengan kesehatan mental; gangguan tidur yang konsisten terbukti meningkatkan risiko gejala depresi secara signifikan melalui gangguan ritme sirkadian [25].
b. Pola Makan: Nutrisi memengaruhi fungsi otak melalui *gut-brain axis*, di mana kebiasaan makan yang tidak sehat (seperti melewatkan sarapan) berkontribusi pada kerentanan emosional [25].
c. Stres Finansial: Kondisi ekonomi yang tidak stabil dan beban keuangan merupakan stresor psikososial kuat yang memicu kecemasan kronis dan memperburuk kondisi depresi pada usia produktif [1].

d. Tekanan Pekerjaan: Paparan terhadap beban kerja yang tinggi, lingkungan kerja yang tidak mendukung, serta tekanan profesional merupakan faktor risiko utama munculnya *occupational depression* [24].
e. Tekanan Akademik: Pada kelompok pelajar dan mahasiswa, tuntutan performa akademik yang intens serta beban tugas yang berlebihan menjadi pemicu utama stres yang berujung pada gejala depresi [27].
f. Umur: Faktor usia menentukan tingkat kerentanan terhadap depresi, di mana transisi biologis dan sosial pada rentang usia tertentu berkaitan dengan akselerasi penuaan biologis yang dipicu oleh stres mental [1], [3].

2.2.2 Klasifikasi dalam Machine Learning

Klasifikasi merupakan teknik *supervised learning* yang bertujuan untuk memprediksi label kategori (target) berdasarkan variabel input (fitur) yang telah tersedia. Dalam penelitian ini, klasifikasi digunakan untuk mengelompokkan individu ke dalam dua kategori biner: "Depresi" atau "Tidak Depresi". Tantangan utama dalam klasifikasi data kesehatan mental adalah menangani sifat data yang bersifat non-linear serta adanya interaksi kompleks antar variabel predictor [9]. Penggunaan algoritma pembelajaran mesin memungkinkan sistem untuk mengenali pola-pola halus tersebut secara otomatis untuk menghasilkan prediksi yang lebih presisi dibandingkan metode diagnostik manual [14].

2.3 Framework/Algoritma yang digunakan

Dalam penelitian ini digunakan pendekatan *data mining* untuk menganalisis dan memprediksi kemungkinan seseorang mengalami depresi berdasarkan beberapa variabel gaya hidup. Algoritma yang digunakan dalam penelitian ini adalah *Random Forest* sebagai *baseline* dan *Random Forest* yang diintegrasikan dengan teknik *oversampling* SMOTE.

2.3.1 Kerangka Kerja CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) adalah kerangka kerja standar yang digunakan sebagai panduan sistematis dalam menjalankan proyek analisis data dan *machine learning*. Metodologi ini terdiri dari enam tahapan yang bersifat iteratif:

1. Business Understanding: Tahap awal ini berfokus pada upaya memahami tujuan dan kebutuhan proyek dari sudut pandang domain masalah yang dihadapi. Fokus utamanya adalah menerjemahkan pengetahuan dan kebutuhan domain tersebut ke dalam definisi masalah *data mining*, serta merancang rencana awal untuk mencapai tujuan yang telah ditetapkan [1], [27].

2. Data Understanding: Tahap ini dimulai dengan proses pengumpulan data awal yang relevan dengan ruang lingkup proyek. Aktivitas di dalamnya meliputi eksplorasi data untuk mengenali karakteristik dasar, mendeteksi masalah kualitas data (seperti adanya data yang hilang atau *outliers*), serta menemukan wawasan menarik pertama yang dapat memicu hipotesis awal [19], [20].

3. Data Preparation: Tahap persiapan data mencakup semua aktivitas yang diperlukan untuk membangun dataset akhir yang siap dimasukkan ke dalam algoritma pemodelan. Proses ini bersifat padat karya dan meliputi pembersihan data (*data cleaning*), transformasi data, seleksi fitur (*feature selection*), penanganan ketidakseimbangan data (*class imbalance*), hingga rekonstruksi variabel baru [19], [20].

4. Modeling: Pada tahap ini, berbagai teknik dan algoritma klasifikasi atau prediksi dipilih dan diterapkan pada dataset yang telah disiapkan. Karena beberapa algoritma memerlukan karakteristik data yang spesifik, proses ini sering kali berjalan beriringan dengan tahap persiapan data untuk melakukan

penyesuaian parameter (*hyperparameter tuning*) demi menghasilkan performa model yang optimal [9], [16].

5. Evaluation: Setelah satu atau beberapa model dibangun dan diuji secara teknis, tahap evaluasi dilakukan untuk menilai kualitas model secara menyeluruh sebelum masuk ke fase implementasi. Fokus utamanya adalah memastikan bahwa kinerja model tidak hanya unggul pada metrik performa statistik (seperti akurasi atau sensitivitas), melainkan juga benar-benar mampu menjawab dan menyelesaikan masalah yang telah didefinisikan pada tahap awal [11], [15].
6. Deployment: Tahap akhir dari CRISP-DM adalah pengintegrasian model yang telah tervalidasi ke dalam lingkungan operasional yang nyata. Output dari tahap ini dapat bervariasi tergantung pada kebutuhan proyek, mulai dari pembuatan visualisasi dasbor interaktif, integrasi model ke dalam sistem aplikasi penunjang keputusan, hingga penyusunan laporan analisis komprehensif bagi pemangku kepentingan [20], [31].

2.3.2 Random Forest

Random Forest merupakan salah satu algoritma pembelajaran mesin berbasis *ensemble learning* yang bekerja dengan membangun sekumpulan pohon keputusan (*decision trees*) secara acak selama proses pelatihan. Keunggulan utama algoritma ini terletak pada kemampuannya untuk menangani dataset berdimensi tinggi dan variabel yang memiliki hubungan non-linear yang kompleks [9]. Dalam konteks penelitian kesehatan mental, Random Forest sangat efektif dalam memetakan interaksi tersembunyi antara berbagai prediktor, seperti bagaimana kombinasi antara durasi tidur yang buruk dan tekanan akademik secara simultan memperburuk risiko depresi [20]. Selain itu, melalui mekanisme *bagging* dan *feature randomness*, algoritma ini mampu

meminimalkan risiko *overfitting* serta menghasilkan prediksi yang lebih stabil dan andal dibandingkan dengan model pohon keputusan tunggal [9].

Setelah seluruh pohon keputusan terbentuk secara independen melalui proses pelatihan, klasifikasi akhir untuk data uji ditentukan menggunakan mekanisme suara terbanyak atau *Majority Voting*. Setiap pohon keputusan akan memberikan satu prediksi kelas secara mandiri, dan kelas yang memperoleh akumulasi suara paling dominan dari seluruh pohon akan ditetapkan sebagai output akhir dari model *ensemble*. Fungsi keputusan matematis dari mekanisme *Majority Voting* ini dirumuskan sebagai berikut:

$$Y = \arg max_k \sum_{i=1}^{B} I(T_i(x) = k)$$

Rumus 2. 1 Rumus Random Forest

Di mana merupakan model klasifikasi *ensemble* utama, adalah hasil prediksi kelas dari pohon keputusan individu ke-, menyatakan total pohon keputusan yang dibangun dalam model, melambangkan variabel target biner (0 atau 1), dan merupakan fungsi indikator yang bernilai 1 jika prediksi pohon individu sesuai dengan kelas dan bernilai 0 jika sebaliknya.

Karakteristik utama dari algoritma *Random Forest* terletak pada ketangguhannya (*robustness*) dalam memproses dataset multidimensi yang memiliki hubungan non-linear kompleks serta interaksi rumit antar variabel gaya hidup. Selain meminimalkan risiko terjadinya *overfitting* melalui pendekatan agregasi hasil *voting*, keunggulan lain dari model ini adalah kemampuannya yang inheren untuk mengukur tingkat kontribusi masing-masing variabel prediktor melalui metrik *Feature Importance*. Melalui metrik berbasis penurunan *Gini Impurity* tersebut, faktor-faktor sehari-hari yang paling dominan

memengaruhi risiko depresi dapat diidentifikasi secara terukur dan transparan.

Tahap optimasi dilakukan dengan menggabungkan algoritma *Random Forest* dengan *GridSearchCV* pada dataset yang telah diseimbangkan dengan SMOTE. Penggunaan *Cross-Validation* (cv=3) memastikan bahwa setiap kombinasi parameter diuji pada bagian data yang berbeda untuk menjamin stabilitas performa. Secara spesifik, metrik *scoring='recall'* dipilih sebagai prioritas utama dalam proses optimasi ini. Hal ini dilakukan karena dalam konteks deteksi dini kesehatan mental, kemampuan model untuk meminimalkan kesalahan deteksi pada penderita depresi yang sebenarnya (*false negative*) jauh lebih penting daripada metrik akurasi secara keseluruhan.

Tahap pemodelan menggunakan algoritma *Random Forest* tidak hanya bergantung pada parameter standar (*default*), melainkan memerlukan proses pencarian parameter optimal (*hyperparameter tuning*) untuk menghasilkan performa klasifikasi risiko depresi yang terbaik. Pada penelitian ini, metode yang digunakan untuk mencari kombinasi parameter terbaik adalah *GridSearchCV*. Metode ini bekerja dengan cara menguji seluruh kombinasi *hyperparameter* yang telah ditentukan di dalam sebuah ruang pencarian (*grid search*) secara mekanis dan komprehensif, dibantu dengan teknik *k-fold cross-validation* (dalam hal ini menggunakan) untuk menjamin objektivitas hasil evaluasi pada data latih.

Sebelum proses pencarian dijalankan, ditentukan terlebih dahulu ruang pencarian parameter (*hyperparameter grid*) berdasarkan karakteristik algoritma *Random Forest* dan batasan komputasi. Ruang pencarian parameter yang diuji dalam proses *GridSearchCV* disajikan pada Tabel 4.3

Tabel 2. 2 Ruang Pencarian Hyperparameter Random Forest

Hyperparameter	Deskripsi Parameter	Nilai yang diuji	Parameter terpilih
n_estimators	Jumlah pohon keputusan (*trees*) yang akan dibangun di dalam hutan.	[50, 100, 200]	200
max_depth	Batas kedalaman maksimum untuk setiap pohon keputusan.	[None, 10, 20, 30]	10
criterion	Fungsi untuk mengukur kualitas pemisahan (*split*) pada *node*.	[‘gini’, ‘entropy]	gini

Melalui total kombinasi yang terbentuk dari Tabel 4.3 dan dikalikan dengan 5-*fold cross-validation*, sistem melakukan iterasi secara menyeluruh hingga menemukan satu kombinasi terbaik yang memberikan performa paling optimal berdasarkan metrik evaluasi yang telah ditetapkan. Hasil akhir *GridSearchCV* menetapkan bahwa kombinasi parameter paling ideal untuk memodelkan risiko depresi pada dataset ini adalah n_estimators=200, max_depth=10, dan criterion='gini'.

Hal krusial yang dilakukan dalam konfigurasi *GridSearchCV* pada penelitian ini adalah penetapan argumen scoring='recall'. Secara umum, optimasi model pembelajaran mesin sering kali diarahkan untuk

memaksimalkan nilai akurasi global (*accuracy*) atau *F1-Score*. Namun, dalam konteks klasifikasi medis atau deteksi dini gangguan kesehatan mental seperti depresi, penggunaan akurasi dapat menyesatkan (*misleading*), terutama ketika berhadapan dengan data yang tidak seimbang (*imbalanced data*).

2.3.3 SMOTE (Synthetic Minority Over-sampling Technique)

Synthetic Minority Over-sampling Technique (SMOTE) adalah metode pra-pemrosesan data yang dirancang khusus untuk mengatasi masalah ketidakseimbangan data (*imbalanced data*). Masalah ini sering muncul dalam dataset kesehatan mental, di mana jumlah responden yang mengalami depresi (kelas minoritas) biasanya jauh lebih sedikit dibandingkan dengan kelompok sehat (kelas mayoritas). SMOTE bekerja dengan cara mensintesis sampel baru pada kelas minoritas menggunakan teknik interpolasi antar data yang berdekatan, alih-alih melakukan duplikasi sederhana [19]. Pendekatan ini memungkinkan model klasifikasi untuk mempelajari karakteristik kelompok berisiko tinggi secara lebih mendalam, sehingga mencegah terjadinya bias di mana model cenderung memprediksi semua individu ke dalam kelas mayoritas. Penggunaan teknik ini telah divalidasi mampu meningkatkan sensitivitas model dan akurasi klasifikasi secara signifikan dalam mendeteksi kasus-kasus depresi yang sebenarnya [19].

Prosedur sintesis data baru dilakukan dengan mengukur jarak kedekatan antar sampel pada ruang fitur kelas minoritas, kemudian menciptakan titik data baru di sepanjang garis acak yang menghubungkan sampel acuan dengan tetangga terdekatnya. Persamaan matematis untuk menghasilkan sampel sintetis tersebut dinyatakan sebagai berikut:

$$x\_new = x\_i + rand(0,1)\ x\ (x\_(zi\ ) - x\_i)$$

Rumus 2. 2 Rumus SMOTE

Dimana merepresentasikan data sintetis baru yang dihasilkan, merupakan sampel asli dari kelompok kelas minoritas, adalah vektor posisi dari salah satu tetangga terdekat yang terpilih secara acak dari kelas minoritas tersebut, dan menyatakan bilangan acak dalam rentang nilai antara 0 sampai 1. Melalui mekanisme interpolasi ini, ruang keputusan yang dibentuk oleh model menjadi lebih luas, general, dan realistis tanpa menimbulkan risiko terjadinya kejenuhan model atau *overfitting* yang tinggi pada data uji.

Teknik SMOTE bekerja dengan cara menggenerasikan sampel baru yang bersifat sintetis pada kelas minoritas (kelompok depresi) berdasarkan kedekatan jarak tetangga (*k-nearest neighbors*). Walaupun SMOTE sangat efektif dalam mengatasi ketidakseimbangan kelas, keberadaan data sintetis berpotensi menimbulkan risiko *overfitting* yang tinggi jika arsitektur pohon keputusan dibiarkan tumbuh terlalu dalam tanpa batasan (max_depth=None).

Jika max_depth tidak dibatasi atau diatur terlalu dalam (misalnya max_depth=30), setiap pohon di dalam *Random Forest*akan berusaha memisahkan data hingga mencapai kondisi yang sangat murni (*pure node*). Akibatnya, model akan "menghafal" pola spesifik atau *noise* dari data sintetis hasil generate SMOTE tersebut, bukan mempelajari pola umum dari variabel Sleep Duration, Dietary Habits, Financial Stress, Age, dan Daily Pressure. Ketika model yang *overfitting* ini diuji menggunakan data uji (X_test), performanya akan turun drastis karena model kehilangan kemampuan generalisasi.

Dengan terpilihnya max_depth=10, model dipaksa untuk melakukan pemangkasan (*pruning*) secara tidak langsung. Batasan kedalaman ini membuat pohon keputusan berhenti melakukan pembelahan sebelum strukturnya menjadi terlalu spesifik. Hal ini menjaga agar model tetap fokus pada tren makro data dan mengabaikan variasi *noise* kecil dari data SMOTE.

Selanjutnya, terpilihnya parameter n_estimators=200 mengonfirmasi bahwa penambahan jumlah pohon hingga 200 unit mampu meningkatkan performa *ensemble*. Prinsip dasar *Random Forest* adalah menggabungkan prediksi dari banyak pohon keputusan independen melalui mekanisme *majority voting*. Semakin banyak pohon yang terlibat (hingga titik jenuh tertentu), varians dari model akan semakin menurun, sehingga prediksi yang dihasilkan menjadi lebih stabil dan tidak sensitif terhadap fluktuasi data individual. Sementara itu, penggunaan criterion='gini' (Gini Impurity) dipilih oleh algoritma karena mampu memberikan efisiensi komputasi yang lebih baik dibandingkan entropy dalam mengukur heterogenitas data pada setiap pembelahan cabang, tanpa mengurangi akurasi hasil klasifikasi risiko depresi.

2.4 Transparansi (Feature Importance)

Meskipun Random Forest sering dikategorikan sebagai *Black Box*, algoritma ini menyediakan metrik *Feature Importance*. Hal ini memungkinkan peneliti untuk mengurutkan variabel mana yang memberikan kontribusi paling besar terhadap prediksi kondisi depresi [12].

2.5 Tools/software yang digunakan

Dalam penelitian ini digunakan beberapa tools dan software untuk membantu proses pengolahan data, analisis, serta pembuatan model prediksi. Tools yang digunakan antara lain:

1. Python

   Python dipilih sebagai bahasa pemrograman utama karena fleksibilitasnya dalam menangani tugas-tugas pengolahan data yang kompleks secara efisien. Bahasa ini didukung oleh ekosistem komunitas yang sangat luas serta ketersediaan ribuan pustaka (*library*) khusus untuk bidang *machine learning* dan ilmu data. Penggunaan Python memungkinkan peneliti untuk mengimplementasikan algoritma tingkat

tinggi dengan sintaks yang ringkas dan mudah dipahami. Hal ini sangat krusial dalam memastikan replikabilitas dari seluruh tahapan pemodelan risiko depresi yang dilakukan dalam penelitian ini.

2. Jupyter Notebook

Jupyter Notebook digunakan sebagai lingkungan pengembangan terintegrasi (*Integrated Development Environment*) untuk menulis, menguji, dan mendokumentasikan kode secara interaktif. Platform ini memungkinkan peneliti untuk membagi proses pengolahan data ke dalam blok-blok kode yang terpisah sehingga memudahkan dalam tahap eksperimen model secara bertahap. Selain itu, kemampuannya dalam menampilkan visualisasi secara langsung di dalam dokumen sangat membantu dalam melakukan inspeksi data secara cepat. Penggunaan perangkat ini memastikan setiap langkah dalam alur CRISP-DM terdokumentasi dengan baik dan mudah untuk ditinjau kembali.

3. Library Scikit-Learn (Sklearn)

Scikit-Learn merupakan pustaka utama dalam ekosistem Python yang digunakan untuk mengimplementasikan berbagai algoritma *machine learning* yang telah terstandarisasi. Pustaka ini menyediakan modul-modul siap pakai untuk proses klasifikasi, termasuk algoritma *Random Forest* yang menjadi fokus utama dalam penelitian ini. Selain tahap pemodelan, Scikit-Learn juga berperan penting dalam tahap evaluasi melalui fitur *confusion matrix*, kurva ROC, serta perhitungan skor presisi dan akurasi. Keberadaan pustaka ini menjamin bahwa implementasi algoritma dilakukan dengan metodologi yang tervalidasi secara matematis.

4. Pandas & NumPy

Pandas dan NumPy adalah dua pustaka fundamental yang berfungsi sebagai tulang punggung dalam manajemen data di dalam lingkungan

Python. Pandas digunakan secara intensif untuk memuat dataset dari sumber eksternal, melakukan pembersihan data (*data cleaning*), serta manipulasi struktur data dalam bentuk tabel (*DataFrames*). Sementara itu, NumPy menyediakan fungsi-fungsi operasi matriks dan komputasi numerik berperforma tinggi yang diperlukan untuk pengolahan fitur-fitur penelitian. Sinergi antara kedua pustaka ini memungkinkan peneliti untuk mengubah data mentah menjadi format yang siap untuk diproses oleh algoritma pembelajaran mesin secara optimal.

5. Matplotlib & Seaborn

Visualisasi data dilakukan dengan memanfaatkan pustaka Matplotlib dan Seaborn untuk menghasilkan representasi grafis yang informatif dan berkualitas tinggi. Matplotlib berfungsi untuk membangun fondasi visualisasi dasar, sementara Seaborn digunakan untuk menciptakan grafik statistik yang lebih estetis dan mudah untuk diinterpretasikan. Pustaka ini digunakan untuk menampilkan berbagai hasil analisis, seperti grafik korelasi antar variabel, perbandingan akurasi model, hingga plot *feature importance*. Melalui visualisasi yang tepat, hasil temuan mengenai faktor pemicu depresi dapat dikomunikasikan secara lebih jelas dan objektif.